MACHINE LEARNING
FOR ASTROPHYSICS
2ND EDITION CATANIA, 8-12 JULY, 2024

Book of Abstracts

Contents

Monday, 8 July 2024

Past and future multiwavelength all-sky surveys / 1

Expediting the discovery process: Application of unsupervised machine learning algorithms to multi-wavelength astronomical datasets

Invited Speaker: Dalya Baron (Carnegie Observatories, USA)

Throughout history, new observations of the night sky have led to new discoveries, often leading to paradigm shifts in astronomy. These have been made possible by observing new wavelength ranges; mapping smaller scales with an increasing spatial resolution; or peering deeper into the more distant universe; to name a few. Over the past decades, astronomy has been going through a data revolution, with numerous surveys mapping astronomical sources from radio to X-ray, and in some wavebands, also as a function of time. Although the discovery space has increased exponentially, trends and groups of interest may be more challenging to uncover due to the increased data complexity. However, such information-rich datasets offer a new opportunity: to use the data itself to form novel hypotheses, a core approach in the field of data science. This talk will focus on one particular type of complex astronomical datasets – multi-wavelength datasets fused from different large-scale surveys. I will start the talk by describing the major transformation astronomy has undergone in that respect, highlighting the challenges astronomers face when analyzing these heterogeneous and high-dimensional datasets. I will then describe how unsupervised machine learning algorithms can be used to reduce the dimensionality and visualize the data; identify simple trends or groups that extend throughout the different survey observables; and detect outliers that may represent completely new types of objects; and by that, expedite the discovery process.

Past and future multiwavelength all-sky surveys / 2

Deblending and photometry of faint sources using VAEs and multi-wavelength data

Speaker: Fernando Caro (Istituto Nazionale di Astrofisica)

The next generation of wide-field surveys is set to deliver an impressive amount of imaging data during the upcoming years when projects such as Euclid or the Rubin Observatory enter in their regular operation phases. The expected improvements on the depth and the magnitude limit of the images that will be acquired pose new challenges related to the increase in the number of blended sources. In this context, efficient mechanisms to deblend sources and then provide photometric estimates are crucial for the case of faint sources since they tend to be ignored by traditional algorithms when they are in the vicinity of brighter sources. To deal with this situation, we have trained several variational autoencoders using simulations that consider different scenarios (faint source located next to a bright star, a bright galaxy, another faint source, etc) which has allowed us to recover multiple faint sources that are typically

lost in the processing of traditional pipelines. We will discuss the results derived from the testing of these VAEs, which have been designed to deblend and measure fluxes, for the case of Euclid and Rubin, based purely on simulated data, while, for the case of HST and JWST, we will discuss the results obtained using real observations. The analysis of these results has also allowed us to draw multiple conclusions regarding transfer learning and the use of a machine trained using the data and simulations from one survey to process data from another one.

Past and future multiwavelength all-sky surveys / 3

Measuring the intracluster light fraction using machine learning

Speaker: Louisa Canepa (University of New South Wales)

The intracluster light (ICL) is an important tracer of a galaxy cluster's history and past interactions. Up to now, only small samples of ICL have been studied due to its very low surface brightness, the difficulties in comparing measurement methods, and the high level of manual involvement needed for most measurements. To process the amount of data expected from upcoming imaging surveys like Euclid and the Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST), we need fully automated and fast methods to make these measurements. In this talk, I will present our work developing a machine learning model that can predict the ICL fraction in cluster images from the Hyper Suprime-Cam Subaru Strategic Program. I will show that this model has been able to effectively learn this task even with very few real data samples, using transfer learning from ICL fractions in artificially generated images. I will describe the methods we use that could be useful for similar machine learning applications in astronomy, and discuss the potential applications of this model with particular focus on upcoming large surveys like LSST.

Past and future multiwavelength all-sky surveys / 4

Machine Learning-Driven Determination of 3D Structures of Molecular Clouds in High-Resolution mm Images

Speaker: Efe Öztaban (Sabancı University)

The analysis of dust scattering halos (DSH), caused by the intense radiation from Galactic black hole transients (GBHTs), provides observational insights to understand the composition and structure of molecular clouds in the line of sight to the source and determine the distance to the source. Observations of DSH of 4U 1630-47, a GBHT, identified a significant molecular cloud (MC -79) that shapes the observed DSH as well as the additional clouds in the line of sight, as evidenced by Chandra and Swift data. The structural analysis of the halo provided a kinematic distance estimate, yet the 'near-far' distance ambiguity for most clouds persists, necessitating future multi-wavelength observations (Kalemci et al. 2018).

Unraveling the near-far distance ambiguity for molecular clouds present in the DSH of 4U 1630-47 requires the determination of the detailed 3D morphology of molecular clouds in the line of sight by utilizing high-resolution 12CO J = 1 - 0 maps and spectral data acquired from the APEX.

We have developed a novel methodology with an application of a machine learning algorithm for clustering for identifying the three-dimensional structures of molecular clouds from sub-mm wave images. Our method consists of four core stages: finding local peaks through Gaussian decomposition, dividing the data into segments, applying OTSU thresholding to identify

cloud-occupied regions, and clustering the data into clouds with a modified version of the Mean-Shift (MS) clustering algorithm. The development of this novel methodology was driven by the necessity of analyzing data with a narrow field of view coupled with the complexity of cloud overlaps within the observational data. Overlapping clouds in the data require an approach to assigning one pixel to more than one cloud by applying a pixel division strategy. This strategy is achieved by the modification of the MS algorithm which allows the creation of soft clustering assignments and division of the pixels with the soft assignment results for the overlapping areas.

Through our approach, we have identified the 3D structures of 15 molecular clouds along the line of sight of the source object, providing pivotal data to resolve the distance ambiguity associated with these clouds. This innovative approach holds potential for broader applications in analyzing various DSHs and molecular clouds within the interstellar medium.

Past and future multiwavelength all-sky surveys / 5

A domain-adaptation approach to classify ionised nebulae in nearby galaxies

Speaker: Francesco Belfiore (INAF - Arcetri Astrophysical Observatory)

The classification and the determination of physical parameters for astrophysical objects is often performed by comparing data with theoretical models. This comparison, however, can be biased by imperfect matching of the observational signatures in the data, and by simplifying assumptions and missing physics in the models. These factors limit the performance of classical machine learning tools and the reliability of uncertainty estimations.

In this talk I will introduce a domain-adversarial neural network (DANN) to bridge the gap between theoretical models (source domain) and observational data (target domain). A DANN is an example of domain-adaptation algorithms, whose goal is to maximise the performance of a model trained on labelled data in the source domain on an unlabelled target domain by extracting domain-invariant features.

I will showcase an application to the classification of ionized nebulae in large integral field spectroscopy datasets. Classifying ionised nebulae in nearby galaxies is crucial to study stellar feedback mechanisms and understand the physical conditions of the interstellar medium.

Our results indicate a significant improvement in classification performance in the target domain when employing the DANN framework compared to a classical neural network classifier. The combined use of domain adaptation and noise injection improves the classification accuracy in the target domain by 20%. This study highlights the potential of domain adaptation methods in tackling the domain shift challenge when using theoretical models to train ML pipelines in astronomy.

Past and future multiwavelength all-sky surveys / 6

A semi-supervised “cluster-then-label” scheme for photometric classification of evolved stars

Speaker: Cristobal Bordiu (INAF-Osservatorio Astrofisico di Catania)

Evolved stars are the main drivers of Galactic evolution from a chemo-dynamical perspective. Therefore, a comprehensive census of these sources is essential for the study of stellar feedback. However, identifying evolved stars is a time-consuming task that often involves spectroscopic observations. With the advent of next-generation telescopes (Roman, LSST) capable of

delivering deep, all-sky photometric surveys at an unprecedented scale, the need for automated, photometry-based methods becomes apparent.

Deep learning offers an opportunity to leverage the vast amount of data available in astronomical archives. In particular, supervised classification techniques have proven to be an efficient solution for multiple science cases, though suffer from the need for (1) large annotated datasets (often incomplete, unreliable and time-expensive); and (2) from intrinsic class imbalances, that reflect the varied sizes of the different stellar populations.

In this talk, we propose a "cluster-then-label" scheme to mitigate these problems. Initially, we employ a small annotated dataset to explore the underlying structure of the data space using density-based clustering. Subsequently, new unlabelled data points are projected into this space and assigned labels based on their location relative to clusters of labelled objects. The annotated data points serve as a ground truth to inform a classifier, which then identifies the optimal decision boundaries to separate meaningful groups. We will present the preliminary results of this work, which has the potential to spot new evolved star candidates within a sample of unclassified sources.

Past and future multiwavelength all-sky surveys / 7

Machine learning for star parametrization

Speaker: Alessio Turchi (INAF-OAA)

An increasing amount of photometric data are made available by multiple large photometric surveys such as GAIA, SDSS, SKYMAPPER, among others. While the large statistical samples available from these catalogs are fundamental for astronomical analyses, parameters estimates (such as temperature, metallicity, or gravity) are difficult to obtain due to the limited reliability of analytical methods applied to photometric data. The Survey of Surveys (Tsantaki et al. 2022) makes a huge contribution in homogenizing high- and intermediate-resolution spectroscopic surveys like APOGEE, GALAH, Gaia-ESO, LAMOST, and RAVE, and provides one of the largest catalogs of radial velocities to date. With similar methods, it also provides a catalog of stellar parameters for a few million stars with uncertainties of about 150 K in temperature, 0.2 dex in surface gravity, and 0.1 dex in metallicity. We made use of this huge dataset of homogenized measurements to train our machine learning algorithms to compute temperature, metallicity, and gravity from photometric data and thus provide spectroscopic-quality estimates for the above quantities on a large sample of stars. At the moment, we in fact obtain performances that are very similar to the uncertainties of the spectroscopic catalog used for the training. We aim to develop a tool that can be applied to the largest possible collection of photometric catalogs in order to map the largest sample of stars in our galaxy.

Past and future multiwavelength all-sky surveys / 8

LEMON: Lens modeling with Bayesian neural networks

Speaker: Valerio Busillo (Istituto Nazionale di Astrofisica)

Strong gravitational lensing is a powerful tool for constraining the total mass and mass density profiles of galaxies, by exploiting visible geometrical features such as lensed arcs, rings and multiple images around massive lens galaxies. Upcoming stage IV surveys, such as Euclid, are expected to discover between 10^4 and 10^5 new gravitational lenses over a total of 14,000 square degrees of the sky, far exceeding the current observed sample. However, traditional modeling of these lenses is often a slow and computationally expensive process. This hinders our

ability to fully exploit these lenses for scientific purposes. To address this challenge, we have developed LEns MOdeling with Neural networks (LEMON), a Bayesian Neural Network which can estimate parameters of galaxy-scale gravitational lenses, with associated uncertainties, in around 9 $ms/lens$. We trained LEMON on a large suite of simulated Euclid-like strong lenses, modeled as singular isothermal ellipsoids plus companions. We find that LEMON can recover all mass and light profile parameters of these lenses, with an accuracy comparable to traditional lens modeling techniques. we finally used LEMON to predict the parameters of "Euclidized" lenses, which have been modified to match the expected data quality of the Euclid survey, and real Euclid lenses. This allows us to check if LEMON can generalize its performance to real lenses, paving the way for the scientific exploitation of the upcoming strong lens discoveries.

Past and future multiwavelength all-sky surveys / 9

Conditional Invertible Neural Networks for enhanced analysis of young, low-mass stars in Trumpler 14: Classical vs. Deep learning methodologies

Speaker: Da Eun Kang (University of Bologna)

Low-mass stars account for the majority of the stars formed in star-forming regions and about half of the total stellar mass. Living longer than massive stars, low-mass stars still remain in the early phases of stellar evolution even when massive stars are dead and provide important information for studying stellar evolution and planet formation. The spectral classification of stars is the most fundamental step in understanding stellar systems. Correctly characterizing the stellar parameters is essential because it significantly influences the interpretation of stellar ages and masses on the H-R diagram and further analysis of the stellar system. In this talk, we will present our recent works on a conditional invertible neural network (cINN) as a novel method to estimate stellar parameters (effective temperature, surface gravity, extinction, and veiling factor) of young, low-mass stars from their optical spectra. The advantage of our network is not only the time-efficient nature of the machine learning technique but also that it can overcome instrumental limitations (low spectral resolution) and methodological limitations of the classical classification method. After pretesting the applicability of cINN in the study of low-mass stars by using template stars, we utilise our tool to analyse about 2000 young, low-mass stars in Trumpler 14 in the Carina Nebula Complex, observed with VLT/MUSE. We compare stellar parameters measured by our cINN with literature values measured from the same data by using the classical template fitting method. We will also talk about the degeneracy between stellar parameters and the importance of considering veiling on the stellar classification.

Past and future multiwavelength all-sky surveys / 10

Exploring the effects of external photoevaporation in Trumpler 14, with the use of neural networks

Speaker: Katia Gkimisi (University of Bologna)

Stellar clusters often host massive stars interacting with the circumstellar material around low-mass stars. The large amounts of far-ultraviolet radiation (FUV) produced by these massive stars cause the effect of external photoevaporation, where gas from the outer layers of the disk around a low-mass star gets depleted. It has been shown that this process could directly affect the evolution of the circumstellar disk and, therefore, planet formation. We have investigated the effects of external photoevaporation in disks around young stars in Trumpler

14. This region contains more than 20 OB-type stars and shows very low extinction, making it a perfect target to investigate the effect. Using MUSE/VLT optical Integral Field Spectroscopic data, we obtained low-resolution optical spectra for several thousand stars in the cluster. We used a conditional invertible neural network (cINN) developed by Kang et al. (2023) to classify the young low-mass stars and estimate their photospheric parameters. The network was trained using synthetic spectra, together with some well-characterized observed spectra from class III template stars. In its initial form, the network takes as input the spectrum of a star and derives the effective temperature (Teff), the gravity (log g), the extinction (AV), and the veiling (rveil). We have shown that this technique effectively classifies young low-mass stars, but the classification becomes challenging as the stellar effective temperature increases above ∼4200K. We will report on the results of our investigation of improving the network's classification methodology, including a more accurate treatment of the effect of veiling.

Past and future multiwavelength all-sky surveys / 11

From Data to Discovery: AI in Radio Astronomy

Invited Speaker: Tao An (Shanghai Astronomical Observatory)

The field of radio astronomy is entering a new era, driven by the Square Kilometre Array (SKA) and its precursors. As we are in the Big Data era, the integration of machine learning (ML) and deep learning (DL) techniques has become crucial to address the complex challenges posed by large and complicated datasets. This talk will explore the transformative potential of AI in radio astronomy, focusing on specific examples from SKA precursor surveys and data challenges. Key areas of focus will include data processing, source detection and classification, and predictive analysis and discovery of unkonwns. Notable examples from SKA precursor surveys, such as the Australian SKA Pathfinder (ASKAP), will be presented to illustrate how AI is improving the processing and analysis of astronomical data. In addition, the talk will address the data challenges of the SKA, emphasising its role in fostering innovation and collaboration within the scientific community. Finally, we will outlook on the future of AI in radio astronomy, discussing ongoing research, potential advances, and upcoming challenges. This talk aims to inspire further exploration and collaboration in this exciting interdisciplinary field.

Past and future multiwavelength all-sky surveys / 12

Applying Self-Organising Maps to identify and classify complex radio morphologies in next-generation radio surveys

Speaker: Afrida Alam (University of Hull)

Radio galaxies often exhibit complex morphologies that are challenging to identify using traditional source finders, necessitating visual inspection of radio maps. The large data volumes of the current and next generation of radio continuum surveys makes such visual inspection impractical. The objective of my research is to develop a more efficient method to identify and classify complex radio sources, including potentially rare objects such as Odd Radio Circles (ORCs). I am utilising an unsupervised machine learning method, specifically a Self-Organising Map (SOM), which has been trained on 251,259 sources from the Rapid ASKAP Continuum Survey (RACS). Morphological labels (compact, extended, double, and triple sources, among others) were assigned to the SOM neurons through visual inspection. These labels were then transferred to the sources in the catalogue via their best-matching neurons. Visual inspection of a sample of our catalogue showed that the reliability of the SOM-derived morphological labels in our catalogue exceeds 90% for approximately 80% of

our sources, demonstrating its efficacy in identifying different types of radio sources. Notably, our method identified approximately 84,424 potential double sources with a >90% reliability. The trained SOM can be used to predict whether new radio galaxy images can be classified as complex sources or not. This project has the potential to improve the classification of complex radio sources, including those with rare or unusual morphologies, in future surveys. This talk will delve into detailed results and implications and discuss the potential impact of our approach on large-scale radio surveys.

Past and future multiwavelength all-sky surveys / 13

Radio Galaxy Zoo: EMU - paving the way for EMU cataloguing using AI and citizen science

Speaker: Hongming Tang (Department of Astronomy, Tsinghua University)

The Evolutionary Map of the Universe (EMU) survey is an ongoing large-scale radio continuum survey conducted by ASKAP, which will discover around 40 million radio sources. While conventional source-finding algorithms could handle ∼90% of source cataloguing in EMU, there might be 4 million sources with well-extended and complex morphological structures awaited to be identified through visual inspection or reliable machine-learning methods. In this talk, I will introduce the Radio Galaxy Zoo: EMU (RGZ-EMU) citizen science project, explaining how the combined usage of citizen science and machine learning helps to characterise the observed radio emission of these complex sources and associate them with belonging host galaxies we use multi-wavelength observations of the same sky area. This especially applies to our radio source tag system generated via natural language processing methods and sample selection processes with the aid of image complexity measurement algorithms. International team efforts in outreach and education will also be mentioned.

Past and future multiwavelength all-sky surveys / 14

WALLABY: HI source-finding with a machine learning framework

Speaker: Li Wang (CSIRO)

The data volumes generated by the WALLABY atomic Hydrogen (HI) survey using the Australian SKA Pathfinder (ASKAP) is necessitating greater automation and reliable automation in the task of source-finding and cataloguing. To this end, we introduce and explore a novel deep learning framework for detecting low Signal-to-Noise Ratio (SNR) HI sources in an automated fashion. Specifically, our proposed method provides an automated process for separating true HI detections from false positives when used in combination with the Source Finding Application (SoFiA) output candidate catalogues. Leveraging the spatial and depth capabilities of 3D Convolutional Neural Networks (CNNs), our method is specifically designed to recognize patterns and features in three-dimensional space, making it uniquely suited for rejecting false positive sources in low SNR scenarios generated by conventional linear methods. As a result, our approach is significantly more accurate in source detection and results in considerably fewer false detections compared to previous linear statistics-based source finding algorithms. Performance tests using mock galaxies injected into real ASKAP data cubes reveal our method's capability to achieve near-100 completeness and reliability at a relatively low integrated SNR∼3 – 5. An at-scale version of this tool will greatly maximise the science output from the upcoming widefield HI surveys.

Past and future multiwavelength all-sky surveys / 15

A fully convolutional neural network to detect diffuse sources in radio surveys

Speaker: Chiara Stuardi (INAF - Istituto di Radioastronomia)

The forthcoming generation of radio telescope arrays promises significant advancements in sensitivity and resolution, enabling the identification and characterization of many new faint and diffuse radio sources. Conventional manual cataloging methodologies are anticipated to be insufficient in fully exploiting the capabilities of new radio surveys, necessitating the integration of artificial intelligence. The analysis of radio interferometric images presents a novel challenge for image segmentation tasks due to random noise and artifacts characterized by diverse statistical properties. Additionally, the conventional workflow involves computationally expensive processes, such as a tailored subtraction of point-like sources, which is particularly challenging for large datasets.

In response to these challenges, we introduce Radio-UNet, a fully convolutional neural network based on the U-Net architecture. Radio-UNet is designed to detect faint and extended sources in radio surveys. Notably, even if it was initially developed for a distinct scientific application, our deep learning model achieves high accuracy in detecting diffuse radio sources while circumventing issues related to noise, artifacts, and point-like sources. Radio-UNet was trained on synthetic radio observations built upon cosmological simulations. To validate the efficacy of our approach, we applied Radio-UNet to a sample of galaxy clusters, where the detection of diffuse radio sources conventionally relied on customized data reduction and visual inspection.

Our results establish the applicability of Radio-UNet to extensive radio survey datasets, probing its efficiency on cutting-edge high-performance computing (HPC) systems. This pioneering approach represents an advancement in optimizing the exploitation of forthcoming radio telescope arrays and leveraging their capabilities for scientific exploration.

Tuesday, 9 July 2024

Past and future multiwavelength all-sky surveys / 16

Building Robust and Impactful Machine Learning Models in Astronomy

Invited Speaker: Daniela Huppenkothen (Netherlands Institute for Space Research)

Machine learning has rapidly become a tool of choice for the astronomical community. It is being applied across a wide range of wavelengths and problems, from the classification of transients to neural network emulators of cosmological simulations, and is shifting paradigms about how we generate and report scientific results. At the same time, this class of method comes with its own set of best practices, challenges, and drawbacks. In this talk, I will highlight a set of recommendations we recently devised for implementing and evaluating machine learning research in astronomy in a way that ensures the accuracy of the results, reproducibility of the findings, and usefulness of the method. The rapid pace of development in machine learning research means that these practices and challenges will likely be in flux for years to come. While we have collected what we consider to be key features of a thoughtful, nuanced approach to machine learning in astronomy, I envision this talk as a starting point for further discussion and would like to invite the participants of this meeting into a conversation about these recommendations, where they fall short and where they might evolve in the future.

Past and future multiwavelength all-sky surveys / 17

Neural network for improving the flux density estimations in polarization of compact sources

Speaker: Laura Bonavera (ICTEA - Universidad de Oviedo)

The importance of galaxy clusters and extra-galactic sources seen as compact or point sources (PS) in ground- and space-based CMB experiments has been clear since the conception of the Wilkinson Microwave Anisotropy Probe and Planck missions. PS in the microwave regime are mainly blazars that affects recovery of the CMB anisotropy signal especially at small angular scales. Unresolved polarized PS could be one of the main contaminants to the primordial B mode detection. Therefore, it is important to develop highly performing methods for PS detection in polarization. On the other hand, the importance of such estimation is not only important for CMB recovery, but also for the astrophysical studies of extragalactic sources. We are exploring the application of Convolutional Neural Networks to enhance the performance of the flux density estimations. Artificial Neural Networks are artificial intelligence techniques involving numerical mathematical models inspired by the human brain, which can be trained to represent complex physical systems by supervised or unsupervised learning. Some approaches, such as the Convolutional Neural Networks, have been successfully applied to image processing (and related fields) for modelling and forecasting and their evolution. An overview of our

Neural Network approach and the results obtained so far will be presented. We develop and train a machine learning model called POSPEN to learn how to estimate the polarisation flux density and angle of point sources embedded in cosmic microwave background images knowing only their positions. We studied the performance of our model with the detected sources of the Second Planck Catalogue of Compact Sources (PCCS2) at the frequencies with polarization measurement and typically improving the number of estimations, especially in the 217 GHz Planck map which is the poorest one among the PCCS2 catalogue.

Past and future multiwavelength all-sky surveys / 18

Deep learning tools for detecting and segmenting galactic structures and contaminants in low surface brightness images

Speaker: Adeline Paiement (LIS lab, Université de Toulon)

The diffuse tidal stellar features that surround galaxies bring a testimony on past galaxy collisions. Because of their low surface brightness and the presence of multiple image contaminants, they are particularly difficult to isolate with classical IA techniques. We present new deep learning techniques to finely characterise the morphology of galaxies from large datasets of deep optical imaging. They are designed to address the specific challenges of these astrophysics data (dynamic range, contrast, noise and contaminants). In particular, we adapt the architecture of a neural network to improve its sensitivity to both short and large scale oriented textures, in order to locate dust clouds contaminants that are overlaid with the imaged galaxies. We also introduce an adaptive scaling layer that allows the neural network learning to focus on relevant parts of the dynamic range in order to reveal the low brightness structures. Our method is trained on a dataset of finely segmented galactic structures, which is built using a home-made annotation tool and a semi-automatic dataset augmentation strategy.

Past and future multiwavelength all-sky surveys / 19

Detection of Galaxy Tidal Features using Self-Supervised Machine Learning

Speaker: Alice Desmons (University of New South Wales)

Low surface brightness substructures around galaxies are a valuable tool in the detection of past or ongoing galaxy mergers. The properties of these substructures, along with the properties of their host galaxy, can tell us about a galaxy's past interactions and the evolution of galaxy populations. In order to draw accurate and statistically robust conclusions about this evolution process, we require a large sample of galaxies exhibiting tidal features. In this talk I will present promising results from a Self-Supervised Machine Learning Algorithm, trained on Hyper Suprime-Cam Subaru Strategic Program (HSC-SSP) data, that can be used to automate the detection of large samples of galaxies possessing low surface-brightness tidal features. Automating tidal feature detection will drastically increase the speed at which large samples can be assembled. I will describe the methods involved in isolating these galaxies and how the same methods can be applied to larger future surveys like Rubin Observatory's Legacy Survey of Space and Time. The algorithm has been applied to $\sim$37,000 galaxies drawn from the HSC-SSP Ultradeep survey. I will present our analysis of this tidal feature sample.

Past and future multiwavelength all-sky surveys / 20

The GaXNet tools for Stage-IV Surveys and Cosmology with Galaxies and Galaxy Clusters

Speaker: Nicola Napolitano (University of Naples Federico II)

The upcoming all-sky surveys from ground and space will collect detailed imaging and spectroscopical information for up to billions of galaxies. These huge datasets encode fundamental information related to cosmology and galaxy formation mechanisms. We have an unprecedented chance to fully exploit galaxies as laboratories for the cosmology and the physics of dark and baryonic matter (i.e. stars and gas) simultaneously. We present a series of deep learning tools to measure the most significant physical parameters for galaxies targeted by imaging and spectroscopical stage-IV surveys (e.g. LSST, Euclid, 4MOST) and novel techniques to fully exploit the multi-dimensional parameter space including all scaling relations among the measured parameters (both for galaxies and cluster of galaxies). We have started to test these techniques on the current stage-III survey (e.g. KiDS and SDSS) for the feature extraction, while we have forecasts for cosmological inferences using galaxy clusters.

Past and future multiwavelength all-sky surveys / 21

Exploring Multi-Band Imaging for Identifying z>6.5 Quasars: A Contrastive Learning Approach Using HSC Data

Speaker: Laura Natalia Martinez Ramirez (Max-Planck-Institut für Astronomie, Pontificia Universidad Catolica de Chile)

Quasars at z>6 are powerful laboratories to study the growth and evolution of supermassive black holes and massive galaxies, the properties of the intergalactic medium, and the formation of large-scale structures within the first Gyr from the Big Bang. Although these distant objects are the most luminous non-transient sources in the universe, it is challenging to find them because they are scarce (< 1 per Gpc^3 at $z > 6$). The last years have seen a significant increase in quasar demographics at z∼6, but the number of quasars at z>6.5 is still low (and only three quasars known at z>7.5) and biased to the brightest sources. Here, we capitalize on the current large-area sky surveys and recently demonstrated unsupervised learning capabilities to perform a novel search of quasars at z>6.5 by applying a Simple Framework for Contrastive Learning of Visual Representations (SimCLR) to multi-band optical imaging data from the Hyper Suprime-Cam (HSC). With this convolutional neural network based method combined with Uniform Manifold Approximation and Projection (UMAP), we derive a latent representation of the data which enables the selection of a sample of z>6.5 QSO candidates. A QSO evolutionary track, as well as brown dwarfs and stars clusters were identified in the latent space. The extension of our methodology to future surveys like Euclid, opens up a wealth of prospects to probe the high-redshift universe and underscores the importance of machine learning-driven approaches in expanding traditional astronomical techniques for identifying these rare and distant astrophysical objects, while addressing challenges posed by small and biased training datasets.

Past and future multiwavelength all-sky surveys / 22

Predictive Modeling of Galaxy Spectra: A Comprehensive Framework Using Transformer Architecture

Speaker: Ginés Martínez Solaeche (Instituto de Astrofísica de Andalucía)

In this study, we developed a foundational model leveraging transformer architecture to analyze galaxy spectra across optical and UV photometric bands. Our objective is to create a comprehensive model capable of simultaneously predicting the characteristics of stellar populations, emission lines, and photometric redshifts from a given observational dataset. Utilizing data from the SDSS and the DESI survey, we generated simulated observations compatible with the J-PAS. Subsequently, we augmented our dataset with UV emission from the GALEX survey in the FUV and NUV bands. The dataset encompasses a wide range of physical properties for each galaxy, including emission lines in the optical range, stellar mass, metallicity, stellar extinction, and spectroscopic redshift.

Similar to large language models, our transformer is trained in an unsupervised manner, aiming to predict the next element in a sequence based on its predecessors. This approach diverges from conventional deep learning algorithms that typically map a set of inputs (observations) to outputs (physical properties). In our transformer model, any piece of galaxy information can serve as either input or output during training, offering unparalleled flexibility. This method does not require complete observations and measurements for each galaxy; rather, it uses subsets of available data alternately as inputs and outputs in iterative training phases. For instance, in one iteration, the model might predict the EW of Hα using only the blue portion of the spectrum, while in another, the mass of the galaxy and the [OIII] emission line might predict the red spectrum part.

The model is further trained on actual observations from miniJPAS to perform domain adaptation, bridging the gap between simulated and observational data. Our results demonstrate the model's effectiveness in characterizing galaxy properties within a unified framework, traditionally achieved through multiple separate codes. Future enhancements will include training with additional data from DESI and J-PAS and possibly extending to photometry in other wavelengths, such as the infrared.

Past and future multiwavelength all-sky surveys / 23

Multi-Task Neural Nets with Monte Carlo Dropout for Spectral Analysis of Galaxies

Speaker: Michele Ginolfi (University of Florence)

The upcoming era of large-scale astronomical surveys, exemplified by instruments like EUCLID and MOONS, demands innovative approaches for rapid and accurate analysis of extensive spectral data. This talk introduces a pioneering deep learning tool that employs a multi-task convolutional neural network with residual learning to simultaneously derive key physical properties of galaxies, such as stellar mass, star formation rate, metallicity, and redshift, from their spectra. Our approach efficiently encodes spectral information into a latent space, employing distinct branches for each physical attribute, thereby harnessing the power of multi-task learning. To address the crucial aspect of uncertainty in predictions, we incorporate Monte Carlo Dropout, enabling the generation of probability distribution functions for each property by executing multiple inferences. A key feature of our methodology is its interpretability. By examining the latent space within the embedding layers, we demonstrate the model's capability to learn and distinguish important but unlabeled galaxy characteristics, such as differentiating between star-forming and quenched galaxies and identifying galaxy morphologies. We demonstrate preliminary results using simulated data of the MOONS instrument, which will be soon operating at the VLT telescope, showcasing the model's efficacy in accurately predicting redshift and physical properties for high-redshift galaxies, even at low-brightness regimes where traditional methods often struggle. Our model, thus, emerges as a powerful solution for the upcoming challenges in observational astronomy, combining precision, interpretability, and efficiency, crucial for the analysis of the massive datasets expected from next-generation instruments.

Past and future multiwavelength all-sky surveys / 24

Galaxy Morphological Classification via Unsupervised Machine Learning in the Big Data Era led by JWST, EUCLID, LSST and SKA

Speaker: Ilin Lazar (University of Hertfordshire)

The morphological properties of galaxies are important tracers of the physical processes, e.g. minor/major mergers, gas accretion and tidal interactions, that have shaped their evolution. Forthcoming 'Big data' surveys (e.g. LSST/SKA), which will produce exabyte volumes of data will be the new 'normal' in this decade. These volumes will make morphological classification using traditional methods (e.g. direct visual inspection) impractical. Even semi-automated techniques, e.g. supervised machine learning with training sets built via visual inspection, may be difficult, because of the time-consuming nature of creating the training sets. However, unsupervised machine learning, does not require training sets, making it ideal for galaxy morphological classification for large surveys. We present an unsupervised machine learning algorithm, that utilizes hierarchical clustering and growing neural gas networks to group together survey image patches with similar visual properties, followed by a clustering of objects (e.g. galaxies) that are reconstructed from these patches. We implement the algorithm on the Deep layer of the Hyper Suprime-Cam Subaru-Strategic-Program, to reduce a population of hundreds of thousands of galaxies to a small number (∼100) of morphological clusters, which exhibit high purity. These clusters can then be rapidly benchmarked via visual inspection and classified by morphological type showing general morphologies (e.g. elliptical, spirals) but also peculiar/rare objects (e.g. mergers, clumpy discs). Using these morphological clusters, we successfully reproduce many known trends of galaxy properties (e.g. stellar-mass functions, rest-frame colours) as a function of morphological type. Furthermore, we find a rare type of elliptical galaxy in the dwarf regime ($10^8 < M < 10^{9.5} M_{SUN}$) which exhibits the same colors and SFR activity as spirals but resembles elliptical shapes and Sersic profiles (blue elliptical) which we discuss in Lazar et al. (2023). Due to its excellent ability to produce fine morphological classifications of even rare objects with minimal human intervention, in forthcoming papers we will use this algorithm on other future surveys as well (e.g. Euclid, JWST) which will give us the chance to explore in detail the morphological evolution of galaxies as a function of redshift, stellar mass and local environment across 80 per cent of cosmic time from an unprecedented statistical perspective.

Time domain / 1

The ALeRCE broker: a community broker for the Vera C. Rubin Observatory

Invited Speaker: Francisco Forster (Universidad de Chile, Chile)

A new time domain ecosystem is developing thanks to a new generation of large aperture and large field of view telescopes, notably the Vera C. Rubin Observatory. Among the tools required are fast machine learning aided discovery and classification algorithms, interoperable tools to allow for an effective communication with the community and follow-up telescopes, and new models and tools to extract the most physical knowledge from these observations. In this talk I will review the challenges and progress of building the Automatic Learning for the Rapid Classification of Events (ALeRCE) astronomical alert broker. ALeRCE (http://alerce.science/) is a broker that annotates, classifies and provides access to a living database of variable astronomical objects since 2019. ALeRCE is focused around three scientific cases: transients, variable stars and active galactic nuclei, and has become the 3rd group to report most transient candidates to the Transient Name Server. I will also discuss some of the results based on the real-time ingestion and classification of the Zwicky Transient Facility (ZTF) alert stream, from the Asteroid Terrestrial-impact Last Alert System (ATLAS) telescopes,

and the classification of simulated data for the Vera C. Rubin Observatory in the context of the ELAsTiCC challenge.

Time domain / 2

Using bowties to catch radio emitting transients

Speaker: Sergio Belmonte Díaz (University of Manchester)

Short timescale radio-transient events provide a signature of extreme astrophysical phenomena. They are often associated with radio emitting neutron stars and Fast Radio Bursts and have been theorised to be emitted by merging black holes and neutron stars. However, discovering these single pulses is computationally challenging and often does not use all available information when selecting candidates. Modern telescopes generate vast numbers of candidates due to the search parameter space spanning large numbers of beams, dispersion measures and pulse widths. Traditionally, the searches are done on dedispersed time series using thresholding techniques based on the statistical properties of the data. More recently, post processing of the potential sources has generated diagnostic plots that have been passed to machine learning algorithms for improved identification of candidates to be viewed by eye. We have investigated a method which bypasses the thresholding step and proceeds straight to generating images of the dispersion measure-time domain. A Mask RCNN model is then used to perform an instance segmentation task on these images to localise potential new transient sources. The deep learning model returns a classification score, a bounding box, and a binary mask of the candidate. The network was trained to identify sources based on the characteristic bowtie shape of the signal-to-noise degradation expected for a real dispersed pulse. I will present the methodology and the results of the training and application to some real datasets.

Time domain / 3

Bridging the Gap between Spectra and Light-Curves: Exploring Multimodal Learning for the Identification of Broad Absorption Line Quasars

Speaker: Nicolás Guerra-Varas (Università di Roma Tor Vergata, University of Belgrade, INAF-OAR)

Broad-absorption line quasars (BAL QSOs) are those that present strong absorption troughs in their spectra. These are associated with powerful winds and outflows that can reach widths of $\geq$ 2000 km/s and velocities of $\geq$ 0.1c. These outflows can trigger significant active galactic nuclei (AGN) feedback processes, which play a crucial role in the evolution of their host galaxies. So far, the identification of clean BAL QSO samples has involved visual inspection of their individual spectra. However, with forthcoming big data surveys such as the Legacy Survey of Space and Time (LSST), we must develop efficient data science algorithms that are able to identify sources of interest. In this work, we explore the use of multimodal machine learning to combine spectral and time-domain data, with the aim of bridging the gap between what has been needed so far to identify BAL QSOs, i.e. spectra, and what will be available in all-sky time-domain surveys like the LSST, i.e. photometry and light-curves. I will present our initial results obtained by ensembling two separate classifiers: one using time-domain features extracted from light-curves, and the other using a lower-dimensional representation of spectra. Finally, I will comment on some future prospects for multimodal learning applications in astronomy.

Time domain / 4

Unsupervised Machine Learning Techniques for Young Stellar Object Light Curve Characterization

Speaker: Máté Madarász (Konkoly Observatory)

Analysing light curve data presents unique challenges, notably the variability in data lengths, which complicates the use of traditional methods and many machine learning techniques designed for fixed-sized input vectors. In this study, we address these challenges by attempting to characterise Young Stellar Object (YSO) candidate light curves from the Gaia Photometric Science Alerts. The characterisation of YSO behaviour is essential to understand the underlying physics through the initial phases of star- and planet formation. While numerous statistics are available to describe temporal data, these are often less useful in the case of YSOs. Our approach leverages a Long Short-Term Memory (LSTM) based network, complemented by experiments with alternative architectures such as Transformers, to accommodate the variable length of our dataset. Central to our methodology is the implementation of an LSTM-based autoencoder designed to learn a compressed, fixed-size latent space representation of the light curves. This allows for the accurate reconstruction of time series from a uniform-dimensional embedded space, effectively converting variable-length data into fixed-length vector representations. To enhance our analysis, we combined the latent space generated by the autoencoder with statistical features for application in various clustering techniques, including k-means and density-based clustering methods such as DBSCAN. Subsequently, the downprojection of high-dimensional data to a 2-dimensional space using t-SNE was performed, enabling the identification of distinct YSO behaviour patterns. By using our method, we are able to identify a range of variability, including stochastic variability across different amplitudes, single or multiple outliers such as flares, dimming, and outbursting events on various timescales, as well as monotonic rising or fading.

Time domain / 5

Improving ML training sets with Active Learning in Fink

Speaker: Anais Möller (Swinburne University of Technology)

Current surveys are detecting thousands of supernovae (SNe) but only a small percentage is able to be classified using spectroscopy. To harness their full power, many surveys are now classifying SNe using ML algorithms. However, these algorithms rely on training sets which are usually built from spectroscopically classified SNe. In this talk, I will present an Active Learning recommendation system built to improve training sets for early type Ia SNe classification. This system is currently deployed in the Fink broker and is applied to ZTF data. I will present how our recommendation system can select the most promising candidates for spectroscopic follow-up and how our evolving training set improves the algorithm performance. I will conclude with a projection of the impact of this system in the era of the Vera C. Rubin Observatory LSST.

Time domain / 6

A new Deep Learning Model for Gamma-Ray Bursts' light curves simulation

Speaker: Riccardo Falco (INAF/OAS Bologna)

In Multi-Messanger astronomy the rapid and precise detection of transient events such as Gamma-Ray Bursts (GRBs) is fundamental to alert the scientific community of new transients and allow them to perform follow-up observations. For this reason, many space mission teams are developing new detection algorithms for GRBs using classical and machine-learning technologies. To train or test these algorithms, it is necessary to have a large GRB dataset, but lack of data is a challenge. GRBs are rare events, and the available datasets are usually very limited in number. In addition, the existing datasets often only cover some of the different characteristics of GRBs with a balanced population. It therefore becomes essential to have a system to simulate GRBs.

This work aims to develop a Deep Learning (DL) based model for generating synthetic GRBs that closely replicate the properties and distribution of real ones. The DL model was trained to provide a larger set of artificial data that can be used for data augmentation and to develop detection algorithms both with classic techniques and machine learning methods.

We propose a new method to generate GRBs based on a 1D version of VAEGAN (Variational Autoencoder Generative Adversarial Network) based on Convolutional Neural Networks. To create the training dataset we extracted the light curves (LCs) of GRBs presented in the Fourth Fermi-GBM Catalog using only long GRBs (formal separation of T90 at about 2 seconds). To create our dataset, an LC was selected for each detector that detected the GRB as a data augmentation technique. Subsequently, a pre-processing phase was applied filtering only the best LCs resulting in 5964 LCs. After evaluating the distribution of the GRB duration (through the t90 parameter), we set the time series length at 220.

A quantitative analysis comparing the histogram of real LC rates (i.e. 1000 LCs never seen by the model) with the histogram of the synthetic rates (i.e. 1000 generated LCs) resulted in an Euclidean norm distance of 3.6 and a Wasserstein distance value of 0.1. The results show that the synthetic LCs are generated with similar properties to the real ones. Furthermore, we are also working on a conditional version of our model using some physical parameters of the GRBs for a more precise generation. This method can generate synthetic LCs for high-energy astronomy projects such as AGILE, CTA and COSI.

Time domain / 7

Signatures to help interpretability of anomalies in astronomical data

Speaker: Emmanuel Gangler (LPCA)

Machine learning is often viewed as a black box when it comes to understanding its output, be it a decision or a score. Automatic anomaly detection is no exception to this rule, and quite often the astronomer is left to independently analyze the data in order to understand why a given event is tagged as an anomaly. Worst, the expert may end up scrutinizing over and over the same kind of rare phenomena which all share a high anomaly score (quite often due to noisy or bad quality data), while missing anomalies of astrophysical interest. In this presentation, I'll introduce the idea of anomaly signature, whose aim is to help the interpretability of anomalies by highlighting which features contributed to the decision. I'll present concrete applications to the search of anomalies in time domain astrophysics within the framework of the SNAD team.

Wednesday, 10 July 2024

Cosmology & Simulations / 1

Cosmology in the machine learning era

Invited Speaker: Francisco Villaescusa (Simons Foundation, NY, USA)

Recent advances in deep learning are triggering a revolution across fields. In this talk, I will discuss how we can apply these techniques to tackle complex problems in cosmology and astrophysics. I will first review how we can improve our understanding of fundamental physics by constraining the value of the cosmological parameters with the highest accuracy. After reviewing the standard methods used to carry out this task I will show how deep learning can largely outperform it. I will then discuss how cosmological structures on small scales contain a wealth of information about the cosmos and how we can retrieve it, taking into account uncertainties from astrophysics phenomena such as feedback from supernovae and active galactic nuclei.

Cosmology & Simulations / 2

Inferring Galaxy Baryonic Properties from IllustrisTNG Dark Matter Merger Trees with Graph Neural Networks

Speaker: Nick Andreadis (Astronomical Observatory, Ghent University)

As more and more data from recent and upcoming large cosmological surveys become available, the need for equally detailed theoretical models of galaxy formation, including large-volume cosmological simulations emerges. The complexity of such simulations, however, implies a great computational cost, which to this day always leads to a compromise in either resolution or size. As a supplement to cosmological simulations, we perform supervised learning to create a mapping between the baryonic and the dark matter (DM) component in high-resolution hydrodynamical simulations. We use the merger trees from TNG100 of the IllustrisTNG simulation suite, that represent the formation history of DM subhalos in the form of graphs, along with Graph Neural Networks (GNNs), which excel in handling datasets that carry such a distinct intrinsic structure as graphs. We train a GNN to infer galaxy baryonic properties, such as stellar and gas mass, given the host subhalo DM properties across the merger tree. We obtain a potent model that is capable of reproducing galaxy populations that are statistically consistent with simulated ones. This powerful tool can generate entire galaxy populations for cosmological volumes: one can run computationally inexpensive, simple DM-only (N-body) simulations and subsequently augment them, by "painting" galaxies on top of the DM subhalos, enabling the construction of Gpc-scale boxes with baryonic physics. In order to tackle the vastly diverse – in size and complexity – graphs that comprise the merger trees, we utilize a specific class of GNNs called Graph Attention Networks (GATs), which also involve additional attention coefficients that capture the importance of certain features in the graphs to determining the desired outputs. These coefficients also provide a degree of interpretability, since they correspond to actual events in the formation history of

the subhalos, and can thus be studied to extract information about the interplay between galaxies and their hosts' assembly history. This method serves not only as a proof of concept, that the model is not just a "black box", rather it learns physically meaningful representations, but also as a scientific tool that can be used to further understand the complex galaxy-halo connection.

Cosmology & Simulations / 3

Extracting optimal information from cosmological surveys with field-level inference and joint analyses

Speaker: Adrian Bayer (Princeton University / Simons Foundation)

Extracting maximal information from upcoming cosmological surveys is a pressing task on the journey to understanding phenomena such as neutrino mass, dark energy, and inflation. This can be achieved by both making advancements in methodology and by carefully combining multiple cosmological datasets.

In the first part of my talk I will discuss methodological advancements to obtain optimal constraints, focussing on field-level inference with differentiable forward modeling. I will first motivate this approach to both reconstruct the initial conditions of the Universe and to obtain cosmological constraints. I will then tackle one of the bottlenecks of this approach – sampling a high-dimensional parameter space – by presenting a novel method, Microcanonical Langevin Monte Carlo. This method is orders of magnitude more efficient than the traditional Hamiltonian Monte Carlo and will enable scaling field-level inference to the regime of upcoming surveys.

I will then discuss combining multiple cosmological datasets to break parameter degeneracies and calibrate systematics. In particular, I will present the HalfDome cosmological simulations, a set of large-volume simulations designed specifically to model the Universe from CMB to LSS for the joint analysis of Stage IV surveys. I will show how these simulations are being used to mitigate systematics, obtain tighter constraints on cosmological parameters, and as a playground for machine learning applications.

Cosmology & Simulations / 4

EMBER: emulating baryons from dark matter-only simulations over cosmic time

Speaker: Mauro Bernardini (University of Zurich)

Galaxy formation is fueled by inflowing gas from the cosmic web. While on cosmological scales the gas distribution traces the dark matter, the relation between gas abundances and matter decouples at galactic scales and transitions into a highly non-linear regime due to the complex interplay of various astrophysical processes.

Hydrodynamical simulations are currently the most accurate tool to model the co-evolution of dark matter and baryons but are expensive, often do not provide enough constraining statistics and do not probe the entire dynamic range of galaxy masses. Thus, fast and accurate models breaking this trade-off are needed to drive progress.

I will introduce the Emulating Baryonic EnRichment (EMBER) framework, a neural network based approach, that can predict high-resolution baryonic fields like HI from low-resolution dark matter simulations. Designed as stochastic emulators, multiple realizations can be sampled for the same dark matter field, highlighting the statistical power of the approach.

The model allows to create cheap mock fields with similar properties to full hydrodynamical simulations of galaxy formation, while operating on a fraction of the computational cost.

I will demonstrate that EMBER can reproduce gas and HI masses of dark matter haloes across 5 orders of magnitude probing beyond the regime of abundance matching models down to dwarf galaxies. I will outline how EMBER can be extended to include velocity information for a more precise dynamical modeling of galactic in- and outflows as well as additional gas fields, like H2, of particular importance to star formation and galaxy evolution.

Cosmology & Simulations / 5

Calibrating Bayesian Tension Statistics with Neural Ratio Estimation

Speaker: Harry Bevins (University of Cambridge)

Quantifying tension between different experimental efforts aiming to constrain the same physical models is essential for validating our understanding of the Universe. A commonly used metric of tension is the ratio, R, of the joint Bayesian evidence to the product of individual evidences for two experimental datasets under some common model. R can be interpreted as a measure of our relative confidence in a dataset given knowledge of another. The statistic has been widely adopted by the community as an appropriately Bayesian way of quantifying tensions, however it has a non-trivial dependence on the prior that is not always accounted for properly. We propose using Neural Ratio Estimators (NREs) to calibrate the prior dependence of the R statistic. We show that the output of an NRE corresponds to the tension statistic between two datasets if the network is trained on simulations of both experiments observables. Such an NRE can then be used to derive the distribution of all possible values of R, within one's model prior, for observations from the two experiments that are consistent. The observed R for the real datasets, derived using Nested Sampling, can then be compared with this distribution to give a prior independent determination of how in tension the experiments truly are. We demonstrate the method with a toy example from 21-cm Cosmology and a joint analysis of Planck and BAO data.

Cosmology & Simulations / 6

Finding analogues of observed galaxy groups and clusters in cosmological simulations with contrastive learning

Speaker: Urmila Chadayammuri (MPIA)

Galaxy groups and clusters are dynamical sites of galaxy evolution and cosmological structure formation. In particular, the intracluster and intragroup media are the hot, volume-filling gas between the galaxies in these deep potential wells. They host signatures of feedback from AGN, non-thermal processes like cosmic ray diffusion and turbulent motion, and mergers with other large structures. The past few years have finally seen suites of cosmological simulations that produce thousands of galaxy clusters, and even more galaxy groups, in realistic cosmic environments and with realistic feedback processes. Matching these to observed systems is, however, an arduous process, because the identifying features are extended and hard to summarise. Contrastive learning is a powerful tool in computer vision, wherein sets of images of one object - augmented by rotation and zooms - are contrasted against sets of images of other objects. The underlying neural network aims to minimise the distance between images in the same set, while maximising the distance to images in a different set. We harness this technique to find analogues observed of galaxy groups and clusters with telescopes like

Chandra, XMM-Newton and eROSITA in the TNG-Cluster suite of simulations, the largest ensemble of high-resolution simulations of galaxy clusters in a cosmological context to date. We present the detected analogues to several iconic galaxy clusters, and show how these analogues can be used for inferring parameters relating to cosmology, plasma physics, and galactic feedback.

Cosmology & Simulations / 7

Application of physics-informed neural networks to neutron star magnetospheres

Speaker: Petros Stefanou (Universidad de Alicante / Universidad de Valencia)

Physics-informed neural networks (PINNs) have gained massive popularity in recent years as competent PDE solvers in a wide variety of physical applications. Combining powerful machine-learning techniques with information about the physical system at hand, they manage to surpass many obstacles that inhibit other classical numerical methods, such as the curse of dimensionality. In addition, they are able to produce new solutions with different physical parameters, boundary conditions or source terms extremely fast once trained, proving to be more efficient in cases where multiple solutions are required. In this work, we employ PINNs for the solution of neutron star magnetospheres. We show that our solver is suitable enough to reproduced any of the well-established results and give insight to new, unexplored families of solutions. Furthermore, we demonstrate that with the correct configuration of the optimization process that takes place during training, PINNs can be competitive with classical numerical methods in terms of accuracy and computational speed, while maintaining an advantage in terms of flexibility and generality.

Cosmology & Simulations / 8

Evaluating Summary Statistics with Mutual Information for Cosmological Inference

Speaker: Ce Sui (Tsinghua University)

The ability to compress observational data and accurately estimate physical parameters relies heavily on informative summary statistics. In this paper, we introduce the use of mutual information (MI) as a means of evaluating the quality of summary statistics in inference tasks. MI can assess the sufficiency of summaries, and provide a quantitative basis for comparison. We propose to estimate MI using the Barber-Agakov lower bound and normalizing flow based variational distributions. To demonstrate the effectiveness of our approach, we conduct a comparative analysis of three summary statistics: the power spectrum, bispectrum, and scattering transform. Our comparison is performed within the context of inferring physical parameters from simulated CMB maps and highly non-Gaussian 21cm mock observations. Our results highlight the ability of our approach to correctly assess the informativeness of different summary statistics, enabling the selection of an optimal set of statistics for inference tasks.

Time Domain / 8

Unsupervised learning for GRBs classification

Speaker: Nicolò Cibrario (INFN and Università di Torino)

Gamma-Ray Bursts (GRBs) are among the most energetic events in the universe. They can last from a few milliseconds to hours, and their multiwavelength and multimessenger observations allow us to probe physics beyond extremes that can be achieved in terrestrial laboratories. While huge progress has been made in the last decades, much remains to be uncovered regarding the origins and underlying mechanisms of GRBs. The categorization of these bursts is a primary objective of the GRBs research, and some machine learning techniques have already been explored for this task. We propose a new fast GRB classification tool based on modern unsupervised deep learning techniques and relying on a set of inputs which have not yet been explored within the machine learning framework, the GRB waterfalls. This set contains the core prompt information relevant for GRB classification, such as duration, temporal variation, pulse structure, spectral hardness and evolution, and how these parameters relate. We used this dataset to train convolutional autoencoders, achieving promising and insightful results regarding GRBs categorization.

Past and future multiwavelength all-sky surveys / 25

A worthwhile detour via artificial spectra: galaxy properties from photometric images with diffusion models

Speaker: Lars Doorenbos (University of Bern)

Modern spectroscopic surveys can only target a small fraction of the vast amount of photometrically cataloged sources in wide-field surveys. Here, we report the development of a machine-learning framework capable of predicting optical galaxy spectra from photometric broad-band images alone. This method draws from the latest advances in diffusion models in combination with a contrastive network. We pass the entire multi-band galaxy images into the architecture to obtain optical spectra. From these, robust values for galaxy properties can be derived with standard population synthesis techniques and Lick indices.

When trained and tested on 64 x 64-pixel images from the last data release of the SDSS survey, the global bimodality of star-forming and quiescent galaxies in photometric space is recovered, as well as a mass-metallicity relation of star-forming galaxies. The comparison between the ""true"" observed SDSS spectra and the artificially created spectra shows good agreement in overall metallicity, age, Dn4000, and E(B-V) values. Photometric redshift estimates of our generative AI can compete with other specialized deep learning techniques. Additionally, we can predict the presence of an AGN up to an accuracy of 82

With our method, scientifically interesting galaxy properties, normally requiring spectroscopic inputs, can be estimated in future data sets from large-scale photometric surveys alone. The creation of artificial spectra can further assist in creating realistic mock catalogs.

Past and future multiwavelength all-sky surveys / 26

The TransformerPayne: Toward a Foundation Model for Stellar Spectra

Speaker: Tomasz Różański (RSAA Australian National University)

Current and upcoming large-scale spectroscopic surveys are delivering an ever-increasing volume of high-quality stellar spectra. This wealth of data presents both opportunities and

challenges in inferring the parameters of stellar atmospheres. Inference initially relied on polynomial interpolation across a regular spectral grid. This approach was later largely replaced by the use of emulators based on neural networks, which offer efficiency and flexibility. However, these often require extensive grids of stellar spectra and tend to reach a plateau in accuracy. To address this challenge, we introduce TransformerPayne: a wavelength-wise neural network-based emulator designed for the accurate and scalable emulation of stellar spectra. TransformerPayne significantly outperforms its predecessors, by achieving a remarkable reduction in mean absolute error, down to 0.001, when emulating normalized flux. This improvement significantly enhances the precision in inferring parameters of stellar atmospheres, as evidenced by refined Cramér-Rao bounds. Our findings demonstrate that TransformerPayne excels particularly in scenarios with spectral grids ranging from 1,000 to 10,000 spectra, achieving a 2 to 10 times improvement in emulation accuracy with respect to the Payne, multilayer perceptron-based model, measured by mean absolute and mean squared errors. Furthermore, by employing a two-stage training strategy - initially pre-training a base model on simplified physics models, then fine-tuning on more realistic spectra - we achieve a tenfold reduction in the sizes of spectral grids while maintaining emulator quality. Our comprehensive experiments not only establish a new benchmark in the emulation of stellar spectra but also highlight TransformerPayne's favorable scaling properties. Summarizing, these improvements facilitate more precise measurements of effective temperatures, surface gravities, and individual abundances, which are of great interest in astrophysics, while using much smaller spectral grids which can then rely on much more accurate and complex numerical models. Finally, the promising scaling properties open a prospect of a foundation model for stellar spectra crafted for astronomical research that can lead to transformative results similar to those in the domains of audio, images, and language.

FlashTalks / 1

Enhancing Xray Binary Analysis through Deep Learning

Speaker: Antonio Tutone (Istituto Nazionale di Astrofisica)

X-ray binaries, systems composed of a star and a compact object, are pivotal to our understanding of astrophysical phenomena. However, the analysis of their X-ray spectra, critical for unlocking the secrets of these celestial bodies, is notoriously computationally intensive. Traditional approaches often struggle to keep pace with the voluminous data generated by observatories, a challenge set to intensify with the advent of next-generation X-ray observatories like XRISM and Athena. This talk introduces a pioneering project that leverages deep learning to revolutionize the evaluation of complex physical models used in X-ray binary analysis. By integrating Bayesian inference within a deep learning framework, our approach significantly reduces computational time without sacrificing accuracy. This breakthrough is particularly timely, ensuring that the astrophysical community can efficiently process and analyze the large datasets expected from upcoming observatories. Our preliminary results demonstrate not only a substantial reduction in evaluation times but also an improvement in the fidelity of model predictions. This advancement is a testament to the potential of deep learning in astrophysical research, offering a scalable solution to the challenges posed by big data in astronomy.

FlashTalks / 2

LightPred - a deep learning model for stellar properties predictions

Speaker: Ilay Kamai (Technion)

Stellar properties play an important role in understanding the evolution of stars, galaxies, and planetary systems. for example, stellar period is essential for understanding stars' age and stellar inclination is important for understanding planetary formation and dynamics. Despite their importance, it is not trivial to measure most of the stellar properties. On the other hand, since the release of "Kepler" mission there has been vast amount of lightcurves for a large sample of stars which opens the door for data-driven analysis. The inference of period and inclination from lightcurves is based on the change in observed flux due to starspots and in general is a challenging task. The period of star can be affected by other periodicities in the system (the spots emergence rate for example) and can change across different latitudes (differential rotation). The most common technics in analyzing period are autocorrelation function and wavelet transform, and recently Reinhold et al reported the largest sample of period calculations from Kepler data (67139 stars) using both methods. The inclination measurements are much more challenging and degenerate with the spot's latitude (As pointed out by Walckovich, Basri and Valenti in 2013). The common methods to calculate inclinations are Rossiter-maclughling effects and Asterosysmology but both can be applied only on a very small subset of, usually planet hosting, stars. In our work, we focus on the development of a deep learning model that learns to predict stellar inclination and stellar period from simulated lightcurves. The model incorporates the most common non-learnable technic for period analysis (autocorrelation function) and a variation of Conformer- a special transformer architecture designed for time series analysis. By that, the model is able to catch both time dependent properties such as period and more global properties like Inclination. We applied our trained model on Kepler dataset and achieved the biggest current sample of period predictions (>80000). In addition, we were able to, at least partially, resolve the degeneracy between inclination and spot's location, and predict stellar inclination for general Kepler stars for the first time. Our model is simple and can be applied for other stellar properties predictions. It also has the potential to learn meaningful insights about the spot's mechanisms themselves.

FlashTalks / 3

Inferring stellar parameters and their uncertainties from high-resolution spectroscopy using conditional Invertible Neural Networks

Speaker: Nils Candebat (INAF-Osservatorio Astrofisico di Arcetri)

Next generation instruments are prompting a revolutionary surge in collecting a wide variety of data, with an unprecedented advancement in our understanding of the Universe. The stellar community is actively adapting to this transformative era, trying to maximise the information obtainable from observational data, in particular stellar spectra. The next generation of multi-object spectrographs and surveys (e.g., SDSS-V: Kollmeier et al. 2017; WEAVE: Jin et al. 2023; MOONS: Gonzalez et al. 2020 and 4MOST: de Jong et al. 2019) have started or will soon start to observe a number of stars more than an order of magnitude larger than the previous ones. The significant shift in data volume requires the development of novel, fast, and precise codes to preprocess the huge flux of data collected every night and to analyse it in a quick but reliable manner, e.g., through supervised machine learning algorithms, such as Neural Networks inference. These algorithms are characterized by enhanced speed, improved precision at lower resolutions [Wang et al., 2023, Guiglion et al., 2024], and reduced dependence on a priori knowledge of physics due to their data-driven nature. Therefore, they are progressively gaining traction within the astrophysical community. However, not all applications of these algorithms succeed in providing the uncertainties on derived parameters [e.g. Huertas-Company and Lanusse, 2023], given the vast number of hyper-parameters, thereby precluding conventional Bayesian analyses.

In this talk, I will introduce OssicoNN, a neural network expressly devised to address new challenges associated with the determination of stellar parameters. At its core, lies the application of the conditional Inversible Neural Network (cINN), an elegant framework developed

by Ardizzone et al. [2019]. Utilizing the flux uncertainty present in the spectra and the properties of cINN, OssicoNN offers an estimation of the overall uncertainty derived during in the inference process, encompassing both the epistemic error associated with the neural network and the aleatoric error intrinsic to the data. I will demonstrate with OssicoNN, the high-performances of a cINN trained on stellar observational data, specifically the GIRAFFE dataset of the Gaia-ESO Survey, to derive effective temperature, surface gravity, metallicity, and various elemental abundances (Aluminum, Magnesium, Calcium, Nickel, Titanium, and Silicon) from stellar spectra.

FlashTalks / 4

GWEEP: A Deep Learning Toolkit for Gravitational Waves Analysis

Speaker: Daniel Tonoiu (Institute of Space Science, Magurele, Romania)

As space exploration continues to expand, machine learning has proven many times to be an important item in our astrophysical data analysis toolbox, very elegantly navigating the challenges of high computational demands and the complexity of multi-dimensional parameter spaces that often occur when trying to process and understand the Universe's messages. In this study, we have successfully demonstrated the use of machine learning techniques for enabling multi-messenger astronomy with gravitational waves. As demonstrated only a few years ago, during the detection of the event GW170817 [1] by LIGO/Virgo, it is of great importance in a gravitational wave experiment to have a rapid mechanism for alerting other observatories about potential events, so that multi-messenger observations could be performed. In this context, we have developed GWEEP toolkit (Gravitational Wave data analysis using DEEp Learning), a neural network based low-latency alert pipeline for a quick and precise gravitational wave event detection and parameter estimation. We have tested our solution on simulated data from the LISA Data Challenge [2], which is part of the LISA [3] consortium. We will present an overall view of the GWEEP pipeline, describing the functionality of each of its components, the methodology, and the results obtained in detecting gravitational waves from LISA-like data. In the future, we aim to develop a generic toolkit of detecting Massive Black Hole Binary (MBHB) events that can be easily customized and adapted for future gravitational wave detectors, using various interchangeable deep learning models (ANNs, CNNs, RNNs, etc). We plan to perform incremental training on large labeled datasets and make fast predictions on new data for a set of fine tuned models.

FlashTalks / 5

Deep Learning-Based Visual-Binary Source Detection in Stamp Images from Multiple Surveys

Speaker: Ilknur Gezer (Konkoly Observatory)

We present a deep-learning approach for detecting visual-binary sources within 2.5 arcsecs in stamp images from various surveys. The aim is to develop a robust model that identifies the presence or absence of visual-binary sources. The process involves several stages including data preparation, model selection, training, evaluation, and deployment. Stamp images from multiple surveys are annotated with binary labels, indicating the presence or absence of the visual-binary sources. A deep learning model, chosen from state-of-the-art object detection architectures, is trained to learn distinctive features of binary sources. Hyperparameter tuning and validation are performed to optimize model performance. The trained model is evaluated on a separate testing dataset to assess its generalization ability. Once deployed,

the model is utilized to detect visual-binary sources in stamp images from different surveys. Post-deployment monitoring ensures the model's effectiveness in real-world scenarios, with ongoing iterations for continuous improvement. Through this approach, we aim to provide a reliable and efficient solution for visual-binary source detection across diverse survey datasets, facilitating scientific research and analysis in astronomy and related fields.

FlashTalks / 6

Early-stopping SOMs as a tool to classify SSOs in space surveys

Speaker: Simone Sacquegna (Università del Salento - Istituto Nazionale di Fisica Nucleare, sezione di Lecce)

Self-organizing maps (SOMs) are unsupervised machine learning techniques that reduce the dimensionality of the original dataset while preserving its topological structure. Their implementation can be represented as a transformation of n data points each with p variables as clusters of similar data on a two-dimensional map. This approach is widely used in astronomy due to the simplicity of the algorithm and the capability of identifying unknown features emerging from the variable space. In particular, we present its implementation as a classifier of stars, Solar System Objects (SSOs) and galaxies in space surveys with large amounts of data such as TESS. We distinguish this approach from the previous studies on the matter by including an early-stopping method. By training the neural network on an input data set and, at the same time, testing its performance on an independent, randomly chosen, control set, it is possible to minimize the risk of falling into an over-fitting state and therefore losing the genericity which is required of this algorithm. As a number of important space surveys will begin their operations in the coming decade, adding to those launched in the last few years, an algorithm capable of swiftly handling large amounts of data and discerning SSOs from galactic or extragalactic sources will be enormously valuable, for the main scientific objectives of the surveys as well as improving our knowledge of objects in the nearest regions of the sky.

FlashTalks / 7

Inter-annotator consensus: optimizing machine learning for astrophysical feature segmentation and classification

Speaker: Renaud Vancoellie (LIS lab, Université de Toulon)

With the arrival of Euclid/LSST and other large-scale surveys we address the automatic detection and segmentation of galactic features from deep sky images. The training of machine learning and deep learning systems requires manual annotations, which tend to present a high variability between annotators. For complex astrophysical features such as low surface brightness collision debris, even expert annotators do not perfectly agree on the features' exact shape and/or nature. To avoid any ambiguity in the learning process, we propose to exploit the confidence information that is carried by the inter-annotator variability. We define a consensus associated with a new dedicated loss function, allowing us to exploit the inter-annotator variability. This loss mitigates learning based on a pixel-basis confidence measure. Annotators may be weighted within this consensus with respect to their expertise. We experiment with various consensus formulas and galactic features to assess the effectiveness of this strategy in improving the learning of a deep neural network. We obtain improved convergence and accuracy for the segmentation of various structures including low brightness galactic collisional debris.

FlashTalks / 8

Automatic Modelling and Object Identification in Radio Astronomy

Speaker: Richard Fuchs (Technical University of Munich)

Building appropriate models is crucial for imaging tasks in many fields, but often challenging due to the richness of the systems. In radio astronomy, for example, wide-field observations can contain various and superposed structures that require different descriptions, such as filaments, point sources or compact objects. This work presents an automatic pipeline that iteratively adapts models for such complex systems in order to improve the reconstructed images. It uses the Bayesian imaging algorithm resolve, which is formulated in the language of information field theory. Starting with a preliminary reconstruction using a simple and flexible model, the pipeline employs deep learning and clustering methods to identify and separate different objects. In a further step these objects are described by adding new building blocks to the model allowing for a component separation in the next reconstruction step. This procedure can be repeated several times for refinement to iteratively improve the overall reconstruction. In addition, the individual components can be modelled at different resolutions allowing to focus on important parts of the emission field without getting computationally too expensive.

FlashTalks / 9

Morphological classification of radio sources in star-forming regions

Speaker: Elena Díaz-Márquez (Institute of Space Sciences (ICE-CSIC))

Radio continuum emission at centimeter wavelengths is found in association with young stellar objects (YSOs) throughout all stages of star formation, from deeply embedded Class 0 protostars to pre-main sequence Class III objects and thus, radio observations provide crucial insights into the formation and evolution of stars. In this work, we present a new approach implementing machine learning techniques for morphology classification of radio sources in star-forming regions, using data obtained from the Very Large Array (VLA). Our study focuses on the development of a model for image classification aimed at determining the morphology of radio sources within these regions. We employ Stochastic Gradient Descent (SGD) as the classifier due to its efficiency and scalability with large datasets. Preliminary results demonstrate promising performance in classifying the morphology of radio sources based on VLA observations. Furthermore, we discuss the ongoing efforts to enhance the model's capabilities by incorporating additional observations with higher angular resolution. As part of our future work, we also expect to integrate more information, such as the radio spectral index, and consider the emission at other wavelengths, since the classification based only on morphology can be ambiguous due to the coexistence of different radio emission mechanisms. The inclusion of more datasets and additional information is expected to improve the model's accuracy and robustness in morphology classification, providing deeper insights into the complex structures within star forming regions.

FlashTalks / 10

Target Selection for a redshift limited survey with Machine Learning

Speaker: Gursharanjit Kaur (Center for Theoretical Physics, PAN)

The WAVES survey of the upcoming multi object spectroscopy facility, 4MOST, has a key science goal of probing the halo mass function to fainter magnitudes as compared to previous surveys like GAMA. This objective, along with constraints on fibre-hour availability, shapes the criteria for selecting galaxies to be observed by WAVES, categorised into two sub-surveys: "wide" and "deep". The surveys will be both flux-limited and redshift limited. The WAVES-wide survey intends to target ≈ 0.9 million galaxies with Z band (central wavelength 0.88 μm) magnitude $Z \leq 21.1$ mag and a redshift $z \leq 0.2$. For the WAVES-deep, the conditions are $Z \leq 21.25$ mag and $z \leq 0.8$. The selection of galaxies with desired completeness for spectroscopic redshift measurements without a prior idea of the redshift is challenging. The general solution to such a problem is the photometric redshift estimation using the magnitude and color measurements of the target. However, such estimates typically have errors of $\sigma z \geq 0.02$ which might be too large for efficient target selection. Our research intends to use machine learning classification to predict the probability of a source falling within the redshift limit as required by the WAVES survey, rather than estimate each object's photometric redshift. This way, each potential target will have a probability attached that it lies below the required redshift threshold, which will help optimise the survey planning. We use supervised machine learning algorithms such as xg boost, our data being ugriZYJHKs broad-band photometry from KiDS+VIKING, as well as ancillary VST+VISTA observations. The redshift label for calibration is derived from an extensive sample of spectroscopic redshifts including wide-angle and deep surveys overlapping with KiDS and the ancillary fields. Our current results indicate that we should be able to achieve the completeness of 90% of the targets with our methodology which is close to the survey success criteria.

FlashTalks / 11

Spatially Variant Point Spread Functions for Bayesian Imaging

Speaker: Vincent Eberle (Max Planck Institute for Astrophysics)

When measuring photon counts from incoming sky fluxes, observatories imprint nuisance effects on the data that must be accurately removed. Some detector effects can be easily inverted, while others are not trivially invertible such as the point spread function and shot noise. Using information field theory and Bayes' theorem, we infer the posterior mean and uncertainty for the sky flux. This involves the use of prior knowledge encoded in a generative model and a precise and differentiable model of the instrument. The spatial variability of the point spread functions as part of the instrument description degrades the resolution of the data as the off-axis angle increases. The approximation of the true instrument point spread function by an interpolated and patched convolution provides a fast and accurate representation as part of a numerical instrument model. By incorporating the spatial variability of the point spread function, far off-axis events can be reliably accounted for, thereby increasing the signal-to-noise ratio. The developed reconstruction method is demonstrated on a series of Chandra X-ray observations of Cassiopeia A.

FlashTalks / 12

Predicting AGN Obscuration with Machine Learning

Speaker: Ross Silver (NASA GSFC)

I present a new method for predicting the line-of-sight column density (NH) values of Active

Galactic Nuclei (AGN) based on mid-infrared (MIR), soft X-ray, and hard X-ray data. We developed a multiple linear regression machine learning algorithm trained with WISE colors, Swift-BAT count rates, soft X-ray hardness ratios, and an MIR-soft X-ray flux ratio. The algorithm was trained off 451 AGN from the Swift-BAT sample with known NH and has the ability to accurately predict NH values for AGN of all levels of obscuration, as evidenced by its Spearman correlation coefficient value of 0.86 and its 75% classification accuracy. Our algorithm significantly reduces the misclassification of unobscured sources (NH<22.5) as heavily obscured or even Compton-thick, compared to previous methods. We applied the algorithm to 487 AGN from the BAT 150 Month catalog with no previously measured NH values. This algorithm will continue to contribute significantly to calculations of NH for large samples of sources, as well as to finding Compton-thick (CT) AGN (NH $\geq 10^{24} cm^{-2}$), thus enabling us to determine the true intrinsic fraction of CT-AGN in the local Universe and their contribution to the cosmic X-ray background.

FlashTalks / 13

Galaxy stellar and total mass estimation using machine learning

Speaker: Jiani Chu (Tsinghua University)

I will talk about a successful proof of concept test on galaxy stellar and total mass estimation using machine learning. Conventional galaxy mass estimation methods suffer from model assumptions and degeneracies. Machine learning, which avoids many of assumption, can be a potential method to predict galaxy properties. In our work, we use a general sample of galaxies from the TNG100 simulation to investigate the ability of multi-branch convolutional neural network (CNN) to predict the central (i.e., within 1-2 effective radii) stellar and total masses, and the stellar mass-to-light ratio M/L. These models take galaxy images and spatially-resolved mean velocity and velocity dispersion maps as inputs. Such CNN-based models can in general break the degeneracy between baryonic and dark matter in the sense that the model can make reliable predictions on the individual contributions of each component. For example, with r-band images and two galaxy kinematic maps as inputs, our model predicting M has a prediction uncertainty of 0.04 dex. Moreover, to investigate which (global) features significantly contribute to the correct predictions of the properties above, we utilize a gradient boosting machine. We find that galaxy luminosity dominates the prediction of all masses in the central regions, with stellar velocity dispersion coming next. We also investigate the main contributing features when predicting stellar and dark matter mass fractions (f, fDM) and the mass-to-light ratio M/L, and discuss the underlying astrophysics.

FlashTalks / 14

ULISSE: A tool for one-shot sky exploration and its application for detection of AGN and galaxy properties

Speaker: Lars Doorenbos (University of Bern)

We present ULISSE (aUtomatic Lightweight Intelligent System for Sky Exploration), a deep learning tool capable of identifying objects that share morphological and photometric properties with a query prototype based on a single image, effectively creating a list of lookalikes. ULISSE performs a similarity search directly using features extracted from the ImageNet dataset and identifies a list of candidates without the need for any time-consuming neural network training, complex analysis of real photometric and spectroscopic data, or deep understanding of the physics of underlying processes. As a result, it can be readily applied to any new dataset.

We applied ULISSE to two tasks: (i) the detection of active galactic nuclei (AGN) and (ii) the identification of galaxies sharing similar star-formation rates (SFR) and stellar masses. To evaluate our method, we use galaxies from the Sloan Digital Sky Survey with available object classifications (AGN and host galaxy morphology) and spectroscopic estimates of SFR and stellar mass. We find that ULISSE is able to retrieve AGNs with an efficiency ranging between 21% and 65% depending on the prototype (compared to the random-guessing baseline of 12%). We find it to be most effective at retrieving AGN in early-type host galaxies, as opposed to prototypes with spiral- or late-type properties. For the second task, between 65% and 83% of the lookalikes retrieved by ULISSE are within 1dex of the SFR and stellar mass of the prototype, regardless of its morphological type. This improves to between 89% and 99% when considering the separate analysis for SFR and stellar mass parameters, respectively. Based on these results, ULISSE is a promising tool for rapidly retrieving different types of astrophysical objects in current and future wide-field surveys (e.g., Euclid, LSST) that target millions of sources every single night.

FlashTalks / 15

Bayesian Generative Strong Lensing with LensCharm

Speaker: Matteo Guardiani (Max Planck Institute for Astrophysics)

Strong gravitational lensing offers a unique view into the distant cosmos by probing directly the distribution of dark matter and providing independent constraints on the Hubble constant. These research objectives call for the utmost precision in the estimation of the distributions of the lens mass and the source surface brightness. Recent strides in telescope technology promise to provide an abundance of yet undiscovered strong lenses, presenting observations of unprecedented quality. Realizing the full potential of these advancements hinges on achieving the highest reconstruction fidelity, both for the source and the lens. To this end we introduce LensCharm, a novel Bayesian approach for strong-lensing signal reconstruction. LensCharm offers a generative-model framework that allows to concurrently and non-parametrically reconstruct both the source brightness and the lens mass distribution along with their associated uncertainties. We showcase the distinctive strengths of our approach on multi-wavelength astronomical data.

FlashTalks / 16

Unmasking the Hidden: Anomaly Detection in ASKAP's Monitoring

Speaker: Zhuowei Wang (CSIRO)

This study explores the intersection of astronomy and human-machine collaboration, focusing on their transformative role in our exploration of the universe. We examine the Australian Square Kilometre Array Pathfinder (ASKAP) and its monitoring data's crucial role in advancing astronomical discoveries. However, ASKAP's success in charting unprecedented numbers of galaxies presents the challenge of managing and interpreting the resulting 'data explosion'.

To tackle this challenge, we introduce a novel anomaly detection framework that synergizes the strengths of machine learning and human expertise. Our approach adapts the k-nearest neighbours (k-NN) algorithm to the unique characteristics of ASKAP's datasets, incorporating astronomical domain knowledge to enhance its effectiveness. By leveraging a hierarchical data representation and a custom distance metric informed by expert insights, our k-NN

variant is well-suited to identify anomalies that are scientifically meaningful. In our approach, machines process the data and flag potential anomalies, while human experts play a crucial role in interpreting and explaining the flagged instances. The human component brings domain knowledge, intuition, and the ability to contextualize the anomalies within the broader astronomical framework. This human-provided explanation is essential for determining the scientific significance of the anomalies and guiding further exploration.

The proposed human-machine collaboration combines the power of machine learning with human expertise to efficiently handle and analyse the immense data generated by ASKAP, potentially enabling ground-breaking discoveries that expand our understanding of the universe.

Thursday, 11 July 2024

Cosmology & Simulations / 9

Expediting Astronomical Discovery with Large Language Models: Progress, Challenges, and Future Directions

Invited Speaker: Yuan Sen-Ting (Australian National University and Ohio State University)

The vast and interdisciplinary nature of astronomy, coupled with its open-access ethos, makes it an ideal testbed for exploring the potential of Large Language Models (LLMs) in automating and accelerating scientific discovery. In this talk, we present our recent progress in applying LLMs to tackle real-life astronomy problems. We demonstrate the ability of LLM agents to perform end-to-end research tasks, from data fitting and analysis to iterative strategy improvement and outlier detection, mimicking human intuition and deep literature understanding. However, the cost-effectiveness of closed-source solutions remains a challenge for large-scale applications involving billions of sources. To address this issue, we introduce our ongoing work at AstroMLab on training lightweight, open-source specialized models and our effort to benchmark these models with carefully curated astronomy benchmark datasets. We will also discuss our effort to construct the first LLM-based knowledge graph in astronomy, leveraging citation-reference relations. The open-source specialized LLMs and knowledge graph are expected to guide more efficient strategy searches in autonomous research pipelines. While many challenges lie ahead, we explore the immense potential of scaling up automated inference in astronomy, revolutionizing the way astronomical research is conducted, ultimately accelerating scientific breakthroughs and deepening our understanding of the Universe.

Cosmology & Simulations / 10

Recovering the CMB signal with neural networks

Speaker: Jose Manuel Casas (Instituto de Ciencias y Tecnologías Espaciales de Asturias (ICTEA-Uniovi))

Artificial neural networks are powerful machine learning models that can be trained to learn non-linear behaviors from data. In this talk, we present a new promising methodology for separating the CMB signal from foregrounds in Planck realistic simulations in temperature and polarization (formed by the CMB, Synchrotron and dust Galactic emissions, PS and thermal SZ extragalactic emissions and instrumental noise), using a fully convolutional neural network on patches of the sky. Our methodology involves a set of convolutional blocks that make inference over microwave sky patches at three HFI Planck channels, and a set of deconvolutional blocks that output a single patch with the CMB signal. Once trained, the network is validated against new, similar data not used for training. We compute the TT, EE and BB average power spectra of all the recovered patches and compared the results against the true CMB maps and of the residual patches. In temperature, we reach a difference between input and recovered signal less than 50 +- $\mu k2$ for multipoles up to $l = 4000$. Residuals are

two orders of magnitude in the extragalactic region and one order below the input signal in the Galactic plane. In polarization, we recover E and B modes at sub-degree scales (up to $l = 800$): the E-mode is recovered with a difference between true and recovered CMB of 0.1+-0.3 $\mu k2$, while the B mode is recovered with a difference of 0.002+-0.02 $\mu k2$. Residuals are two orders of magnitude below the input E signal and one order below the B one, although noise levels start to dominate the signal along the B-mode. Moreover, we show that training without noise allows us to recover the signal at $l = 1500$ and that training with a proper foreground model is crucial for recovering the signal.

Cosmology & Simulations / 11

Dynamical Modelling of Galactic Kinematics using Neural Networks

Speaker: David Simon (University of Oxford)

The advent of integral field data has revolutionised the study of galaxy evolution. A key component of this is dynamical modelling methods which have allowed for crucial insights to be made from kinematic data. Despite this importance, most dynamical models make a number of key assumptions which do not hold for real galaxies. At the same time, machine learning methods are becoming increasingly powerful, with many applications appearing in astronomy. These have the potential to be used to both improve existing dynamical modelling methods as well as build new ones. To investigate this, we construct a training set of dynamical models of early-type galaxies using Jeans Anisotropic Modelling (JAM). We then train a neural network on this data using the parameters of JAM and mock photometry as the input. We find that we are able to speed up JAM to a remarkable degree while maintaining accuracy in our models. We conclude with some remarks on the application of this to cosmological hydrodynamical simulations and the prospects of producing the next generation of dynamical models.

Cosmology & Simulations / 12

Surrogate Modeling for Supernova Feedback toward Star-by-Star Simulations of Milky-Way-sized Galaxies

Speaker: Keiya Hirashima (The University of Tokyo)

In recent decades, galaxy simulations have found the interdependence of multiscale gas physics, such as star formation, stellar feedback, inflow/outflow, and so on, by improving the physical models and resolution. Still, so-called sub-grid models, simplified or calibrated to specific summary statistics, have been widely used due to the lack of resolutions and scalability. Even with zoom-in simulations targeting Milky-Way-sized galaxies, the mass resolution remains capped at around 1,000 solar masses (e.g., Applebaum et al. 2021), comparable to the mass of molecular clouds. To overcome the limitations, we are developing a new N-body/SPH code, ASURA-FDPS, to leverage exascale computing (e.g., Fugaku), handle approximately one billion particles, and simulate individual stars and stellar feedback within the galaxy. However, achieving sufficient parallelization efficiency presents challenges. Conventional codes have attempted hierarchical individual time-stepping methods to decrease calculation costs, but the emergence of communication costs hinders scalability beyond one thousand CPU cores. One of the causes is short timescale events localized in tiny regions, such as supernova explosions. In response, we have developed a surrogate model using machine learning to duplicate supernova feedback quickly (Hirashima et al., 2023). This model is trained with three hundred 3D simulations of resolved supernova feedback within isolated giant molecular clouds. When a supernova explodes in a galaxy simulation, the model predicts the physical

quantities 100,000 years ahead on the fly. Our new approach reduces the computational cost to ~ 1 percent compared to directly resolving SN feedback. In the presentation, we report the fidelity and progress of the simulations with our new machine-learning technique.

Cosmology & Simulations / 13

Self-supervised learning for component separation for the extragalactic submillimetre sky

Speaker: Victor Bonjean (FORTH)

We use a new approach based on self-supervised deep learning networks originally applied to transparency separation in order to simultaneously extract the components of the extragalactic submillimeter sky, namely the cosmic microwave background (CMB), the cosmic infrared background (CIB), and the Sunyaev-Zel'dovich (SZ) effect. In this proof-of-concept paper, we test our approach on the WebSky extragalactic simulation maps in a range of frequencies from 93 to 545 GHz, and compare with one of the state-of-the-art traditional methods, MILCA, for the case of SZ. We first visually compare the images, and then statistically analyse the full-sky reconstructed high-resolution maps with power spectra. We study the contamination from other components with cross spectra, and particularly emphasise the correlation between the CIB and the SZ effect and compute SZ fluxes around positions of galaxy clusters. The independent networks learn how to reconstruct the different components with less contamination than MILCA. Although this is tested here in an ideal case (without noise, beams, or foregrounds), this method shows significant potential for application in future experiments such as the Simons Observatory (SO) in combination with the Planck satellite.

Cosmology & Simulations / 14

Simulation-based Supernova Cosmology

Speaker: Konstantin Karchev (SISSA)

The field of supernova Ia (SN Ia) cosmology—which already yielded the groundbreaking discovery of dark energy—is set to be revolutionised in the coming years by an avalanche of new data. Their analysis will require a methodology that can easily scale to the $>10^5$ expected SNae Ia while rigorously accounting for systematics. To this end, I will present a SN Ia analysis framework based on truncated marginal neural ratio estimation (TMNRE)—a variant of simulation-based Bayesian inference (SBI)—that addresses a number of outstanding issues. The use of neural networks (NNs) allows one to analyse raw SN light curves without an expensive fitting stage that derives summary statistics. The framework can derive marginal constraints on the cosmological parameters of interest by efficiently and implicitly marginalising the $O(10^5)$ latent variables. Similarly, TMNRE can be used for quick and principled Bayesian model comparison, addressing e.g. the need for additional standardising variables. Lastly, by employing modern set- or transformer-based NN architectures, SBI can seamlessly incorporate selection effects and non-Ia contamination while simultaneously analysing host information to derive photometric redshift estimates, making it a single-stage-to-cosmology analysis framework.

Cosmology & Simulations / 15

Machine-Learnt Star-formation laws: Symbolic Regression with FIRE-2 galaxies

Speaker: Diane Salim (Rutgers University)

Whilst star formation (SF) in the interstellar medium (ISM) and the physics that govern it are some of the most fundamental mechanisms needed to paint a nuanced understanding of galaxy evolution, attempts to construct closed-form analytic expressions that connect SF and physical variables that have been observed to influence it, such as the density and turbulence properties of surrounding gas, still exhibit substantial intrinsic scatter. In this work we leverage recent advancements in machine learning (ML) and use neural network symbolic regression (SR) techniques to produce the first data-driven, ML-discovered analytic relations for SF using the publicly available FIRE-2 simulation suites, which have no explicit numerical sub-grid recipe for SF. We employ a genetic algorithm-based SR pipeline that assembles analytic functions to model a given dataset called PySR, training it to predict symbolic representations of a model for the star formation rate surface density (ΣSFR) at both 10 mega-years (Myr) and 100 Myr based on extracted variables from FIRE-2 galaxies. These variables include those dominated by small-scale characteristics such as gas surface densities, gas velocity dispersions and surface density of stars, as well as large-scale environmental properties like the dynamical time and the potential of gas. The stochastic nature of short-scale SF is reflected in the functional forms of the equations we find via PySR for ΣSFR at 10 Myr, which are heavily dependent on properties that show more fluctuation on a local scale such as the surface density and velocity dispersion of gas. These results indicate the critical need to consider turbulence and validate "bottom-up" approaches in recipes of SF over short timescales. The equations that PySR finds to describe ΣSFR at 100 Myr exhibit more influence from properties that show more consistency over the entire galaxy such as the dynamical time, pointing to the efficacy of "top-down" analytic models for describing SF when considering longer timescales. Furthermore, the equations found for the longer SFR timescale better capture the intrinsic physical scatter of the data within the Kennicutt-Schmidt plane, indicating that training on longer SFR timescales leads to less overfitting. Future applications of this process include investigation of equations found for observational data and input of such found equations into semi-analytic models of galaxies.

Cosmology & Simulations / 16

Toward a deep learning approach for fast galaxy generation

Speaker: Tanner Sether (Michigan Technological University)

Numerous missions are currently collecting enormous amounts of cosmological data, and the near future will only bring more. Providing theoretical predictions to match with observations is key to constraining cosmological parameters. Recently, simulations have gotten extremely detailed and accurate; however, these simulations are computationally expensive. In order to retain the necessary accuracy and speed to keep up with data from surveys, machine learning proves to be a valuable asset. Here, we utilize convolutional neural networks in conjunction with diffusion models and simulations from the CAMELS project to create mappings from relatively inexpensive dark matter simulations to galaxies from detailed but accurate hydrodynamic simulations. Our techniques outperform existing models, such as the Halo-Occupation Distribution method, both in speed and accuracy.

Cosmology & Simulations / 17

ForSE+: Simulating non-Gaussian CMB foregrounds at 3 arcminutes in a stochastic way basing on GAN

Speaker: Jian Yao (SISSA)

We present ForSE+ package to produce non-Gaussian diffuse Galactic thermal dust emission maps at arcminute angular scales along with the capacity to generate random realizations of small scales. This work is based on ForSE (Foreground Scale Extender) package, which was recently proposed to simulate non-Gaussian small scales of thermal dust emission at 12' using Generative Adversarial Networks (GANs). These new realistic maps will be used in the future to understand the impact of non-Gaussian foregrounds on the measurements of the Cosmic Microwave Background (CMB) signal, in particularly on the lensing reconstruction, de-lensing and the detection of cosmological Gravitational Waves in CMB polarization B-modes. With the input of the large-scale polarization maps from observation, ForSE+ is trained to produce realistic polarized small scales at 3' following the statistical properties, mainly the non-Gaussianity, of observed intensity small scales, which are evaluated through Minkowski functionals. Furthermore, by adding different realization of random component to the large-scale foregrounds, we show that ForSE+ is able to generate small scales in a stochastic way. In both cases the output small scales have a similar level of non-Gaussianity compared with real observation and correct amplitude scaling as a power law.

FlashTalks / 17

Morphology and spatial distribution of high-redshift dust emission using component separation and deep learning

Speaker: Arnab Lahiry (Institutes of Astrophysics/Computer Science - Foundation for Research and Technology - Hellas (FORTH))

Observations and state-of-the-art cosmological simulations show that there is a lack of low-mass as well as massive galaxies with respect to the number of dark matter halos, in comparison to the theoretical predictions by the ΛCDM model. For the local universe, this discrepancy can be explained by studied baryonic processes, but it is still unclear what causes the dearth of galaxies in the early universe (beyond cosmic noon : z>3), as gas and dust had much different physical conditions. Through this PhD thesis, we aim to perform a morphological and kinematic analysis of high-redshift galaxies, focusing on the Hot Dust-Obscured Galaxy (Hot DOG) population to understand the physical dynamics of such systems. Our tentative goals are to use data from simulations as well as observations, and incorporate blind source separation techniques to understand the dynamics of different components of interacting sources at high redshift, coupled with novel deep learning methods to learn about the evolutionary stage in the overall galaxy evolution timeline. This extensive morpho-kinematic study would allow us to explain the respective physical processes better, and close the gap between observations/simulations and theoretical predictions. In this talk I will talk about the premise, goals and progress with my PhD thesis.

FlashTalks / 18

Machine learning for radiometer calibration in global 21cm cosmology

Speaker: Samuel Leeney (University of Cambridge)

In this talk we propose a Physics based AI framework for precise radiometer calibration in

global 21cm cosmology. These experiments aim to study formation of the first stars and galaxies by detecting the faint 21-cm radio emission from neutral hydrogen. This global or sky-averaged signal is predicted to be five orders of magnitude dimmer than the foregrounds. Therefore detection of the signal requires precise calibration of the instrument receiver, which non-trivially amplifies the signals detected by the antenna. Current analytic methods appear insufficient, causing a major bottleneck in all such experiments. Unlike other methods, our receiver calibration approach is expected to be agnostic to in-field variations in temperature and environment and furthermore does not rely on assumptions that certain critical components are impedance matched. For the first time we propose the use of machine learning for calibration of global 21-cm experiments.

Cosmology & Simulations / 19

Symbolic regression and interpretable ML

Invited Speaker: Deaglan Bartlett (Institut d'Astrophysique de Paris (CNRS & Sorbonne Université), France)

The drawback of the more traditional, numerical ML techniques is their opaqueness; it is not always clear what information is being used and how methods trained on (necessarily imperfect) simulations will perform when applied to real-world data. An alternative branch of ML – Symbolic Regression (SR) – has clear advantages in this regard. By searching for simple, analytic descriptions of the data, the benefit of SR algorithms over traditional ML methods is their interpretability and clear extrapolation behaviour when employed on data outside the range of the training set. As such, SR has developed into a vibrant field of research in ML and has been increasingly employed within the fields of astrophysics and cosmology. In this talk I will review the diverse range of methods uses within SR, discuss the benefits and drawbacks of these approaches, and highlight the exciting applications of SR in astrophysics.

Cosmology & Simulations / 20

Accelerated inference with neural networks for CMB and LSS analyses

Speaker: Boris Bolliet (University of Cambridge)

We will present state-of-the-art machine learning based tools that can be interfaced with traditional Boltzmann and sampling codes to accelerate inference analyses of CMB and LSS data. We will show real life example with the CMBxLSS code class_sz used in ACT and SO collaboration pipelines.

Cosmology & Simulations / 21

Globular clusters' dynamical age determination based on machine learning performed on large number of detailed MOCCA simulations

Speaker: Arkadiusz Hypki (Faculty of Mathematics and Computer Science of Adam Mickiewicz University)

MOCCA code is able to perform detailed numerical simulations of globular clusters of any size within a few days (http://moccacode.net). Because of its speed and a close agreement with N-body codes MOCCA code is perfect to perform a grid of simulations for various initial conditions. It is currently beyond the capabilities of any N-body codes. At this moment our MOCCA database consists of over 2500 models for broad range of initial parameters ($\sim$ 200 TBs). To handle such amount of data from numerical simulations I created BEANS - web-based software for interactive distributed data analysis with a clear interface for querying, filtering, aggregating, and plotting data (http://beanscode.net). BEANS software relies on software which already proved its value in the industry like Apache Hadoop (distributed processing), Apache Pig (high level language for Apache Hadoop), Elastic (petabyte scale search engine) and more. Recently, a plugin to BEANS was added, which provides simple interface to train machine learning algorithms and then test their predictions. During the talk I will demonstrate how one can query huge amount of MOCCA numerical simulations with Apache Pig to build a training set with some parameters describing the models (e.g. core half-mass radii, relaxation time, pre/post core collapse phase). Then, using a few different machine learning algorithms, I will show how one can train them and finally test their predictions on the dynamical age of GCs (whether a GC is in pre- or post-collapse phase). The algorithms are trained using relatively easy to compute parameters of globular clusters which in turn could provide an alternative, easy method to determine the dynamical ages of the GCs.

SPACE CoE / 1

The SPACE Centre of Excellence

Speaker: Giuliano Taffoni (Istituto Nazionale di Astrofisica)

High-performance computing (HPC)-based numerical simulations are indispensable in astrophysics and cosmology (A&C), aiding scientific understanding of celestial phenomena. These simulations model complex physical processes underlying observations, crucial for interpretation. Advancements in computational power promise transformative scientific discoveries through larger-scale simulations, especially critical in astrophysics, a field devoid of traditional lab experiments. Exascale computing systems pose a challenge as existing numerical codes aren't tailored for these complex architectures, limiting their potential. Recognising this gap, SPACE's primary aim is to adapt current A&C codes for pre-exascale HPC architectures funded by EuroHPC JU and future systems. This initiative, involving scientists, code developers, HPC experts, and hardware & software manufacturers, aims to redesign eight prominent European A&C HPC codes. The goal is to re-design these codes for efficient utilization of upcoming computing architectures. SPACE's efforts also focus on advancing workflows using machine learning and visualisation building an exascale ready numerical Laboratory for A&C.

SPACE CoE / 2

Interpreting the largest cosmological simulations using Representation Learning

Speaker: Sebastian Trujillo (Heidelberg Institute for Theoretical Studies (HITS))

Numerical simulations are the best approximation to experimental laboratories in astrophysics and cosmology. However, the complexity and richness of their outputs severely limit the interpretability of their predictions. We describe a new assumption-free approach to obtaining scientific insights from cosmological simulations. The method can be applied to today's largest simulations and will be essential to solve the extreme data access, exploration, and analysis

challenges posed by the Exascale computing era. Our tool automatically learns compact representations of simulated galaxies in a low-dimensional space that naturally describes their intrinsic features. The data is seamlessly projected onto this latent space for interactive inspection, visual interpretation, sample selection, and local analysis. We present a working prototype using a Hyperspherical Variational Convolutional Autoencoder trained on the entire sample of simulated galaxies from the IllustrisTNG project. The tool produces an interactive visualization of a "Hubble tuning fork"-style similarity space of simulated galaxies on the surface of a sphere. The hierarchical spherical projection of the data can be extended to arbitrarily large simulations with millions of galaxies.

SPACE CoE / 3

Deep learning for scientific data compression

Speaker: Lubomir Riha, Petr Strakos (IT4Innovations, VSB-TUO)

The increasing power of the available supercomputing infrastructure and forthcoming exascale performance enables rapidly more accurate physical simulations, such as detailed high-fidelity CFD simulations or simulations from astrophysical codes. However, as the accuracy of simulations increases, the requirements for storage space also increase quickly. Results from large-scale transient simulations are typically on the order of tens of terabytes to units of petabytes. Supercomputing systems generally increase computational power much faster than the increase in storage space. The standard user space quota is in the order of tens of terabytes. An important issue is the possibility of compressing extensive 3D volumetric data so that, in some cases, simulation results can be stored, transferred and visualised even on infrastructures with limited storage and memory capacity. We present a modified version of Fourier mapping functions for learning the mapping from 3D coordinates space (x,y,z) to an output space containing, in our use cases, simulation variables. We use dynamic Fourier features mapping to (i) decrease the input size in the learning process and (ii) enable dynamic learning of the optimal distribution of features mapping. Results will be shown in several use cases.

SPACE CoE / 4

High-Performance Unsupervised Machine-Learning in Numerical Simulations and Satellite imaging

Speaker: Francesco Tomba (University of Trieste)

Unsupervised machine-learning (UML) techniques play a crucial role in data analytics: clustering algorithms, in particular, are widely adopted in Astronomy. We ported a UML algorithm, the Advanced Density Peak (ADP) from the original Python code to an optimized C code and parallelized it with OpenMP, speeding it up by a factor of 40. We then applied ADP, which couples extreme sensitivity to small signal-to-noise and limited runtimes, to two astrophysical problems: (i) finding substructures in cosmological numerical simulations and (ii) de-blending images from the newly launched EUCLID satellite. We discuss these applications and review the preliminary results. Furthermore, the constant increase in the size of the datasets makes it already impossible to fit the data from cutting-edge experiments on a single node. That is why we are developing a new high-performance distributed-memory parallelization for ADP using MPI, which we also illustrate and discuss as of general interest in this field.

Other / 1

Bayesian Imaging of the Spatio-Spectral X-Ray Sky

Speaker: Margret Westerkamp (Max Planck Institute for Astrophysics; Ludwig-Maximilians-Universität München A)

The advent of the next generation of instrumentation in astrophysics and elsewhere poses several challenges due to the high-dimensional signals that vary in space, time and energy. These typically have non-trivial correlation structures and are often a mixture of overlapping signal components that need to be separated. In order to facilitate multi-instrument analysis of correlated signals in general, we are developing the Universal Bayesian Imaging Kit (UBIK), a flexible and modular framework for high-fidelity Bayesian imaging. UBIK is designed to address these challenges through the use of information field theory, which allows the consistent application of Bayesian logic to signal reconstruction, so that uncertainties can be estimated. In particular, we use generative models to encode prior knowledge about the signals of interest in order to exploit spatial and spectral correlations and thereby improve their reconstruction from noisy data and enhance the component separation. Here, we show the application of UBIK to the Poisson-noise-affected, latest merged Chandra X-ray data on the supernova remnant SN1006, allowing data sets from different observations to be combined. We present a multi-stage approach in which the spatial reconstruction obtained for a single energy range is used to derive an informed starting point for the full spatio-spectral reconstruction in latent space. This provides a high quality visualisation of the complex features of the remnant without the influence of other sources in the field.

Other / 2

Conditional Variational Autoencoders for the analysis of the Gaia spectrophotometric data of pre-main sequence stars

Speaker: Cristina Paola Marcellino (INAF - Osservatorio Astrosico di Catania)

The characterization of the pre-main-sequence (PMS) stars from the Gaia spectrophotometric data is one of the new objectives for the fourth Gaia Data Release (DR4). The mass accretion taking place through the protostellar disc in PMS stars induces an additional component of continuum radiation excess with respect to what is observed in main-sequence stars. Such a continuum excess is detectable in the wavelength range observed by the low-resolution Gaia BP/RP spectrophotometer (300-1100 nm). The Extended Stellar Parametrizer - Cool Stars (ESP-CS) module of the Gaia astrophysical parameters inference system aims at analyzing the BP/RP spectra in order to determine the stellar and accretion parameters for each source: effective temperature, surface gravity, metallicity, mass, extinction, mass accretion rate and filling factor.

The ESP-CS inference method presented in this talk is based on the Conditional Variational Autoencoders (CVAE). We trained the CVAE on a pre-computed grid of BP/RP model spectra parametrized by the stellar and accretion parameters. Our CVAE are not then data-driven, as they are conditioned on model spectra based on the magnetospheric accretion model. CVAE allow us to embed the information on the stellar and accretion parameters into the encoder and decoder networks. In this way, the latent space represents both the BP/RP model spectra and their associated parameters. Convolutional layers are employed for both the encoder and the decoder architectures. The trained CVAE are employed to infer the stellar and accretion parameters of the PMS stars by passing the observed BP/RP spectra to the decoder network. The variational nature of the method allows us to feed different realizations of the latent space into the decoder as samples from statistical distribution, thus providing posterior samples of the inferred parameters. Our Java implementation allows us to analyze one BP/RP spectrum in about 3 seconds on an intel core i7 laptop.

Other / 3

Recurrent Neural Networks for Adaptive Optics parameters estimation

Speaker: Fabio Rossi (INAF)

In the realm of Adaptive Optics (AO), accurately estimating optical and atmospheric parameters such as Strehl Ratio (SR), Seeing (ε), wind speed (ws), External Scale (L0), and the CN2 stratification is crucial. This is particularly vital when aiming to measure these quantities directly from the Wavefront Sensor (WFS) itself. The WFS provides a true measure that directly impacts image quality, as opposed to relying on external instruments which may observe a different optical path. Traditional analytical methods exist for calculating some of these parameters from WFS telemetry data, yet they often fall short when it comes to measuring certain parameters accurately, necessitating approximations that compromise their reliability. Despite the utility of these conventional methods, the advent of statistical approaches leveraging the vast amounts of data generated by WFS telemetry has emerged as a viable alternative in recent years. In this study, we explore a groundbreaking methodology that employs Recurrent Neural Networks (RNNs) within the broader spectrum of machine learning algorithms. These networks utilize data from the WFS (including CCD frames, slopes, reconstructed modes, and actuators commands) to derive estimates of atmospheric and optical parameters. Our findings, which are initially based on simulated data from the LBT telescope's SOUL instrument using PASSATA software and will later extend to real data analysis, are juxtaposed with the current state-of-the-art methodologies to underscore the potential and efficacy of our approach.

Other / 4

Tracing ongoing quenching in jellyfish galaxies with machine learning techniques: Challenges in full spectral classification

Speaker: Ariel Werle (INAF - Padova)

Jellyfish galaxies display long tails of gas that was removed from their stellar disks through a process known as ram pressure stripping. These objects represent a fast transition phase between star-forming and post-starburst/quiescent galaxies. Using spatially resolved spectroscopic information, one can trace this evolution by identifying quenched regions in jellyfish galaxies and studying their properties and spatial distribution. This has been done through visual inspection in select cases where the relevant spectral features are particularly prominent, but extending this analysis to a statistical sample requires the classification of spatially resolved spectra of varying S/N in a large wavelength range that is coeval between different objects, accounting for the displacement of spectral features caused by redshift and internal kinematics. To this end, we developed a pipeline for the selection of a training set and the full spectral classification of star-forming and quenched galaxy spectra with 774 fluxes between 3680 and 6000 angstroms using a 1D convolutional neural network, which accounts for the displacement of spectral features. Similar results can be obtained with other techniques but require the spectra to be heavily smoothed. The classification is applied to a sample of 21 Jellyfish galaxies, revealing ongoing quenching in 6 objects, which is confirmed through visual inspection and stacking. The quenched regions are generally found to be opposite to the tail, with the exception of one object that displays outside in quenching. We show that the central regions of galaxies with ongoing quenching have enhanced star-formation rates with respect to other jellyfish galaxies and to a control sample, showing how the outside-in stellar population gradients found in post-starburst galaxies are shaped during the stripping process.

Other / 5

Using Deep Learning for spectral classification, redshift predictions and anomaly detection

Speaker: Fucheng Zhong (Sun Yat-sen University / University of Naples Federico II)

We present the Galaxy Spectra Network/GaSNet-II, a supervised multi-network deep learning tool for spectra classification and redshift prediction. GaSNet-II can be trained to identify a customized number of classes and optimize the redshift predictions. Redshift errors are determined via an ensemble/pseudo-Montecarlo test obtained by randomizing the weights of the network-of-networks structure. GaSNet-II achieves 92.4% average classification accuracy over the 13 classes and mean redshift errors of approximately 0.23% for galaxies and 2.1% for quasars.

We also utilize a generative network to reconstruct the input spectrum (GaSNet-III). By comparing the input and reconstructed spectra, the redshift and type of the input can be determined. Additionally, the outlier/anomaly spectrum can be detected. We demonstrate that the majority of galaxy and quasar spectra can be reconstructed resulting in a high-accuracy redshift (3 x 10^{-4} for galaxy and 3 x $10^{-}3$ for QSO).

FlashTalks / 19

Predicting the Ages of Galaxies with an Artificial Neural Network

Speaker: Laura Hunt (University of Hull)

We present a new method of predicting the ages of galaxies using an artificial neural network (ANN) with an evaluation of the methods such that future work can be derived for other properties of galaxies with ANNs. Ideally, we would be able to quickly and accurately determine individual properties of galaxies such as age on a large scale for use in studies. However, we currently rely on time consuming and computationally heavy traditional modelling to estimate many properties at a time. This means smaller studies that only require one or two properties are limited to the galaxies that have already been surveyed and modelled. To circumvent this simple machine learning models such as ANNs could be used to predict one or two properties on large samples of galaxies.

We train our ANN to recognise the patterns between the equivalent widths of spectral indices and the mass-weighted ages of galaxies estimated by the MAGPHYS model in data release 3 (DR3) of the Galaxy and Mass Assembly (GAMA) survey. We provide a discussion of the process of optimising our hyperparameters to increase the accuracy of our predictions with Tensorflow and Keras capabilities to provide an accessible method for future studies to build on.

We quantify the quality of our predictions by calculating the mean squared error as 0.02, mean absolute error as 0.11, R-squared score as 0.53 and presenting prediction uncertainties which show our network performs well. Finally, we describe how the predictions are physically motivated as they follow trends already present in the properties and EWs of the galaxies which shows the network is able to meaningfully derive age from the EWs of galactic spectra.

FlashTalks / 20

Machine learning-based photometric classification of galaxies,

quasars and stars

Speaker: Donato Cascio (University of Palermo)

The vastness of astronomical datasets and the intricate nature of celestial objects create a pressing demand for robust classification techniques. Machine learning emerges as a powerful solution in this context, capable of discerning complex patterns and classifications. The work we will present delves into the application of machine learning techniques for classifying astronomical sources based on photometric data, including normal and emission line galaxies, quasars, and stars. Our analysis is grounded in datasets from the Sloan Digital Sky Survey (SDSS) Data Release 18 (DR18), encompassing approximately 100.000 patterns. We will present a comprehensive investigation, evaluating various classification models such as Random Forest (RF), XGBoost (XGB), Support Vector Machine (SVM), and k-Nearest Neighbors (KNN). Our approach involves an in-depth analysis of the metrics used and a meticulous feature selection process aimed at optimizing figures of merit, particularly mean class accuracy. We explore different classification strategies such as one-vs-one and one-vs-all. Furthermore, to reduce the risk of overfitting or under-utilization of data, k-fold cross-validation is performed. The various configurations explored yield promising results.

FlashTalks / 21

Analysis of Velocity Distribution Functions using the Gaussian Mixture Model

Speaker: Beniamino Sanò (Università di Trento)

Velocity distribution functions (VDFs) measured by the Magnetospheric Multiscale (MMS) mission are complex 3D datasets that can be represented as a superposition of multiple beams (Goldman et al., 2020). Recent papers (Dupuis et al., 2020; De Marco et al., 2023) proposed the use of the Gaussian Mixture Model (GMM) to identify different populations. The method is applied systematically to MMS data in the tail and in the dayside of the Earth's magnetosphere. The conclusion of the analysis is that the GMM is capable of detecting the presence of multiple beams within an overall distribution. The GMM can define reliably the complexity of a measured data-set in terms of the number of optimal beams provided by information theory criteria. The goal of the analysis is to differentiate between simple and more complex regions based on the VDF measured by MMS. In particular, complex shaped electron distributions, thus represented with a greater number of clusters, have been shown to be good indicators for processes of interest such as magnetic reconnection and turbulence (Shuster et al., 2014; Hoshino et al., 2001).

FlashTalks / 22

Probing AGN dynamics through a natural language processor lens

Speaker: Benedict Rouse (Pontificia Universidad Catolica de Chile)

We employ natural language processing algorithms with attention, repurposed to receive QSO spectra to predict unseen spectra, broad lines, and super massive black hole masses. We find that the trained algorithm is able to reproduce with high significance masked broad lines and/or continua in QSO spectra, highlighting an ability to learn from and leverage physical information imprinted amongst the entire spectrum. A key implication is that this information may help to refine physical properties such as single-epoch black hole masses.

We tested the algorithm's ability to directly predict black hole masses, with no spectral fitting or decomposition, finding reasonable success to reproduce single-epoch prescriptions. Finally, the algorithm is able to reconstruct broad UV lines such as C IV and Ly-alpha in the presence of broad absorption lines or intervening IGM. We will discuss the attention mechanism, which allows us to peek inside and probe what information is being used to make the above predictions and several broad future applications that we envision.

Friday, 12 July 2024

Astroparticle Physics / 1

Astroparticle Physics and Deep Learning

Invited Speaker: Jonas Glombitza (Erlangen Centre for Astroparticle Physics)

Algorithms based on machine learning have been extraordinarily successful across many domains, including computer vision, machine translation, engineering, and science. Moreover, in the physical sciences, the importance of machine learning is growing fast, driven by the necessity for precise and efficient algorithms that can effectively handle vast amounts of complex and high-dimensional data. Recently, with the help of these novel algorithms, providing improved reconstructions, new insights into astroparticle physics could be gained, raising the question: Could machine learning even become a new paradigm for data-driven knowledge discovery?

In this contribution, we review the state of machine learning in astroparticle physics after introducing the fundamental concepts. We outline the potential of this emerging technology, illustrate the wide variety of possible applications in the context of astroparticle physics, and review the latest breakthroughs. Finally, we present novel approaches and techniques and discuss future applications and challenges in the field.

Astroparticle Physics / 2

Machine Learning Techniques for Space Calorimeter Experiments

Speaker: Maria Bossa (INFN Napoli)

Space-based experiments for direct detection of high-energy cosmic rays often employ optimized calorimeters, designed to achieve high energy resolution and broad acceptance capabilities. However, the significant volume of data collected demands innovative approaches for analysis and interpretation. In this study, we introduce our efforts to develop an AI algorithm dedicated to classifying electromagnetic and hadronic showers. By leveraging machine learning techniques, our algorithm aims to accurately discern and categorize these fundamental particle interactions. The implementation of our AI model is expected to enhance the efficiency and precision of data analysis within experiments, thereby advancing our understanding of high-energy cosmic ray phenomena.

Astroparticle Physics / 3

Graph neural networks for track reconstruction in space

experiments

Speaker: Federica Cuna (INFN-Bari)

For tracking reconstruction purposes, various models inspired by computer vision applications have been studied, operating on an image-like representation of tracking detector data. Image-based methods encounter challenges in scaling up to realistic space experiment data due to high dimensionality and sparsity. Conversely, geometric deep learning methods such as graph neural networks (GNNs) are well-suited to address reconstruction problems in astroparticle physics. GNNs leverage the inherent graph structure of particle tracking data, where tracker hits are naturally represented as nodes and connections between hits as edges or links. By exploiting this graph representation, GNNs can effectively capture complex patterns and dependencies in the data, making them particularly suitable for tasks such as noise versus signal identification in particle tracking applications. We will present a novel GNN approach for particle reconstruction in space experiments. Specifically, by using the Geant4 toolkit, we simulated the response of 4 scintillating fiber tracking layers, aiming to mimic the setup employed during a beam test at CERN, which utilized a 10 GeV negative pion beam. Our approach involves training and evaluating different node-classifying GNN algorithms to distinguish between noise and signal hits in the simulated detector data. These GNN algorithms utilize techniques such as message passing and graph convolution to iteratively aggregate information from neighboring hits and learn discriminative features for classification tasks. We compare the performances of these GNN algorithms with those of traditional analytical tracking approaches, highlighting the advantages of leveraging deep learning techniques for particle reconstruction in space experiments.

Astroparticle Physics / 4

Machine learning enhancements for Cherenkov telescopes data analysis

Speaker: Ambra Di Piano (Università di Modena e Reggio Emiglia; INAF/OAS Bologna)

The Cherenkov Telescope Array Observatory (CTAO) will provide incredible opportunities for the future of ground-based very-high-energy (VHE) gamma-ray astronomy. To optimise its scientific output, the CTAO will have a Science Alert Generation (SAG) system to reconstruct and analyse observations in real time, as part of the Array Control and Acquisition (ACADA) system. This work aims at implementing deep learning algorithms for the enhancement of standard analysis implemented in real-time, where changes in the observational conditions and performance constraints can have a higher impact on the overall sensitivity of the analysis. We developed two applications of Convolutional Neural Network (CNN) based models, trained offline on 20k simulations. One model consists in an autoencoder that learns the background level of a given observation, and subtracts its contribution from the counts map. The second model consists in a 2-dimensional regressor that extracts hotspots for the localisation of candidate sources in the field of view. Finally, we compare results from the two models with results from the standard techniques. We achieve comparable results, with the advantage of not requiring a priori knowledge on the background model for online inference.

Astroparticle Physics / 5

CHEREN-ZOO: an educational project for automatic routines

Speaker: Angelo Adamo (Istituto Nazionale di Astrofisica)

The main goal of this study is to implement a Citizen Science game designed specifically for children and based on the images that the Cherenkov Telescope Array (CTA) instrument will capture. The proposed objective was to start an active dissemination project ("learning by doing") both for humans - while children will simply play a game, humans interested in this project will take part in a citizen science project on 'gamma ray astronomy - both for machines. Thanks to their cataloguing, the result of the use of pareidolia, the participants in the game will teach the automatic algorithms, already quite efficient in discerning the normal traces of particles and photons, to recognize even the most ambiguous ones. It will therefore mainly be an "educational project for automatic routines" which will make use of the results of experiments conducted through structured interviews. Due to their lack of preconceived ideas, this game will mainly involve children in nursery school and the first two classes of primary school with the aim of building a database of human solutions to some pattern recognition problems that still plague automatic routines who work on selecting light tracks captured by Cherenkov telescopes.

Astroparticle Physics / 6

Extending Cosmic Ray Background in the LiteBIRD Experiment using Generative Adversarial Networks

Speaker: Giovanni Cavallotto (INFN Milano Bicocca)

Cosmic rays (CR) reaching telescope detectors in outer space are known to induce glitches and background noise, as the High-Frequency Telescope (HFT) of the LiteBIRD experiment, designed to measure the B modes polarization of the Cosmic Microwave Background (CMB). The presence of CR noise significantly influenced the Planck experiment, which shared similarities in detector design with LiteBIRD. Detecting the faint CMB signal excluding data affected by CR noise from analysis is impractical. In order to address this challenge, it is imperative to accurately simulate the CR background throughout the duration of LiteBIRD's three-year mission. However, state-of-the-art Monte Carlo simulations of CR background incur significant computational overhead, typically requiring 30 times the simulated period, rendering complete coverage of the mission infeasible. To overcome this limitation, we propose augmenting Monte Carlo simulations with Generative Adversarial Networks (GANs). By leveraging GANs, we can efficiently generate a sufficient number genuine, statistically independent images, unlike traditional noise analysis techniques together with template expansion methods.

Space Weather / 1

The application of the Stack-CNN algorithm for the analysis of Mini-EUSO data

Speaker: Antonio Giulio Coretti (INFN / University of Turin)

Machine learning algorithms are widely recognized as powerful and highly efficient methods in a broad range of classification tasks. In particular, Convolutional Neural Networks (CNN) are specifically suitable for the classification of images. In this work, a novel methodology that makes use of a CNN was tested on data from Mini-EUSO, a space telescope currently on board the International Space Station (ISS). Pointing towards the Earth, Mini-EUSO uses photomultiplier tubes to search for Extreme Energy Cosmic Rays (EECR, $E>10^{21}$ eV). In parallel, the instrument is able to detect a wide variety of other atmospheric phenomena, such as meteors. Leveraging this latter capability, the presented method, called Stack-CNN, is applied for the detection of meteor events. Already employed in previous studies, it makes use of a Stacking procedure according to a trial velocity vector, before applying a CNN to

distinguish the signal from the background. An optimization study on the trial velocity vector pool is conducted, aimed at reducing the computational time of the algorithm. Results indicate that our algorithm performs better than the standard meteor detection algorithm, also retrieving with good accuracy the velocity vector associated with the moving meteor. In addition, being able to detect moving luminous objects, this methodology could also be used to detect space debris illuminated by solar albedo. In this presentation we will report the methodology and the results obtained using the Stack-CNN algorithm.

Space Weather / 2

ML Methods For Space Debris Detection in Bistatic Radar Data

Speaker: Miguel Andrea Zammit (University of Malta, Institute of Space Sciences & Astronomy)

Over the last half-century, the exponential increase of satellite launches has led to a significantly large population of debris objects in Earth's orbit, particularly in the Low Earth Orbit (LEO) and Geostationary Earth Orbit (GEO) regimes. Their subsequent detection and monitoring have thus become ever more pertinent, with facilities such as the BIstatic RAdar for LEo Survey (BIRALES) space debris radar regularly used for detecting new NORAD objects and tracking confirmed bodies in the LEO regime. The current detection pipeline installed at BIRALES employs a Multi-beam streak detection strategy (MSDS) algorithm to identify and segment radar streaks in spectrogram data indicative of debris objects. It achieves 90% recall of the track information at a 2dB signal-to-noise ratio (SNR). However, the MSDS's relatively slow computation time will pose a problem once BIRALES is scaled to a larger array. We aim to leverage the power of machine learning to develop a streak detection and segmentation model able to improve on the performance seen by the MSDS algorithm, both in terms of the recall and false positive rates, as well as being able to maintain a level of solid performance at lower SNR levels. The preliminary tests we will present here focus on applying a shortlist of ML streak segmentation and localisation model architectures on BIRALES observational data and comparing their performance with the MSDS.

Space Weather / 3

Convolutional Neural Networks for Detecting Moving Objects in Wide-Field Surveys

Speaker: Belén Yu Irureta-Goyena (École Polytechnique Fédérale de Lausanne)

When not tracked by the telescope, moving objects can leave trails of light in long-exposure astronomical images as they cross the sky. These objects can either be natural (asteroids) or artificial (satellites and space debris). We discuss their detection using machine-learning methods by analysing the data of two wide-field surveys: the Zwicky Transient Facility (ZTF) and the VLT Survey Telescope (VST). First, as almost-unchanged remnants of the assembly of the Solar System, asteroids can provide us with insights into this formation process. In addition, some of them pass close to the Earth's orbit and threaten to collide with it, causing local or even global damage. For these reasons, it is essential that we achieve a comprehensive census of these objects. The trails left by asteroids need to be discerned from other linear features present in the images, such as cosmic rays or columns of bad pixels, which implies that traditional line-detection algorithms do not suffice. To this end, the near-Earth object (NEO) group of the ZTF developed a pipeline that has been used to discover more than 200 near-Earth asteroids on the ZTF images, which cover an extremely large field of view (47

square degrees per exposure). However, the pipeline reports a high rate of false positives that need to be discarded manually. This talk explores new methods to improve the precision of the current NEOZTF pipeline while maintaining the same levels of recovery rate. In particular, it analyses the application of a convolutional neural network (TernausNet) trained on both past detections and simulated asteroid data. Second, satellites and space debris have increased exponentially over the past decade in the near-Earth environment and, as such, we must monitor their population. In this talk, we discuss the application of a convolutional neural network that includes a Hough Transform block to detect the signature tracks of these objects, which are longer than those of asteroids, in images taken with the VST. Lastly, the developed tools will also be applicable to Euclid, LSST and a range of other satellite and ground-based surveys, setting the stage for fast and automated moving-object detection in the upcoming generation of telescopes.

Solar Physics / 1

Advancing Solar Flare Forecasting: Predicting Solar Flares via a Deep Learning Approach Using Time Series of SDO/HMI Line-of-Sight Magnetograms

Speaker: Elizabeth Doria Rosales (Università di Trento / Università della Calabria)

Solar flares, sudden bursts of electromagnetic energy originating from magnetically active regions on the solar surface, pose significant risks to satellite infrastructure, communication systems, and power grids. Predicting these events is crucial for both physicists and governmental agencies. Recent observations and research have revealed that solar flares are more complex phenomena than previously thought, as they extend from the corona to the lower photosphere, underlining the interconnected nature of the Sun's atmospheric layers and emphasizing the need for a comprehensive understanding of solar flare dynamics to improve space weather forecasting and our ability to protect space-based technology and infrastructure. Conventional approaches to this problem rely on features extracted from line-of-sight (LoS) magnetograms of solar active regions, which have traditionally been linked to increased flare activity. More recent methodologies leverage temporal series of LoS magnetograms to extract as much information as possible from available data, progressing towards a fully automated flare forecasting system. Despite these advancements, recent studies with LoS magnetograms haven't shown significant improvements over previous methods, raising doubts about their effectiveness. We propose a deep learning-based approach to address this challenge. We model the solar flare forecasting problem as a binary time series classification task using line-of-sight (LoS) magnetograms obtained from the Solar Dynamics Observatory's Helioseismic and Magnetic Imager (SDO/HMI). These magnetograms provide crucial data capturing magnetic field variations on the solar surface, which are essential for predicting solar activity. Our primary objective is to distinguish between active regions likely to produce M- or X-class flares within a 24-hour window and those remaining inactive. Such a classifier could be instrumental in the development of an early warning system for solar flares. Our proposed approach consists of two phases: firstly, a Convolutional Neural Network (CNN) autoencoder performs feature extraction from the magnetograms. Secondly, a Long Short-Term Memory (LSTM) binary classifier processes the extracted features to predict flare activity. Our proposed methodology achieves a remarkable 90% test accuracy. It is validated against test case AR HARP 377, demonstrating its efficacy in solar flare prediction.

Solar Physics / 2

Bayes in Space: A Bayesian Deep Learning approach for coronal temperature estimation

Speaker: Nikita Balodhi (Northumbria University, Newcastle)

The outermost layer of the Sun - the Corona, is a region of intense activity and showcases a range of solar phenomena affecting the thermal distribution of its constituting plasma. The study of the temperature distribution across the corona is essential in learning about different heating mechanisms that lead to the strikingly high temperatures reached by the solar corona. The temperature distribution can be estimated using photometric observations taken in multiple bandpasses by imaging surveys like the Atmospheric Imaging Assembly (AIA) onboard the Solar Dynamics Observatory. However each bandpass spans a range of temperatures and hence, cannot be estimated directly through these observations. The temperature distributions at each bandpass can be estimated by inverting the intensity or the irradiance i.e. number of photons hitting the detector through the channel passband. We propose an uncertainty based deep learning approach to generate Differential Emission Measure (DEM) maps from solar images, that contain information of the amount of thermal plasma emitted by the solar corona along a specific line-of-sight (LOS) at a certain temperature. We train a neural network to read the AIA image in multiple optically thin channels and develop their DEM maps across a range of temperature bins as an output. We further introduce an uncertainty in the existing deep learning methods for obtaining the DEM maps from AIA images by incorporating Bayesian techniques like variational dropout and bayes by backdrop into our machine learning model, and discuss how these different Bayesian approaches perform on our given data. We compare our uncertainty incorporated results to analytical estimates obtained by regularised least squares methods for DEM inversion.

Exoplanets / 1

Reducing stellar noise in exoplanet observables using machine learning

Speaker: Manuel Perger (Institut de Ciències de l'Espai (ICE, CSIC) & Institut d'Estudis Espacials de Catalunya (IEEC))

The quest for life beyond the Solar System is one of the most thrilling endeavors of our time. Most of the 5600 exoplanets already detected have not been directly observed; instead, we heavily rely on indirect methods such as the transit or radial velocity technique. These methods are already sensitive to discovering Earth twins or detecting so-called biomarkers during a transit in the planetary atmospheres using space-based telescopes like the JWST or ARIEL. However, they are also highly sensitive to stellar activity manifested in magnetic phenomena such as dark spots, bright faculae, or granulation patterns, which can subsequently diminish, mimic, and even hide planetary signals. We use our state-of-the-art StarSim code to model stellar exoplanet observables, to feed neural networks, and to thereby predict the stellar contributions to exoplanet observables. In a first paper, we were able to reduce this stellar noise down to 4.5% for our model data and 10% for observational data of two test stars. I will present our approach, explain the recent results, and outline its capability to advance the most important issues of modern exoplanet science.

Exoplanets / 2

Applications of autoencoders to exoplanetary transit light curves

Speaker: Lucy Smith (University of Hull)

Identifying transits in light curves has become a major method of detecting exoplanets, and

many deep learning classification models have been applied to exoplanetary transit data. We have chosen to explore the potential of deep learning generative models. We consider the ability of a simple autoencoder and convolutional autoencoder to generate light curves of transiting exoplanets, with the aim to help balance the current data received from the recent NASA missions such as Kepler and TESS, as due to the rarity of transiting planets there is a severe data imbalance in observed data (too many non-transits vs transits). We show that a simple convolutional autoencoder can be used to augment transit light curves, which can then be added to the total data set and aid in the classification of exoplanets and non-exoplanets. Finally, we explore a further application by examining the possibility of using a simple autoencoder to classify transits and non-transits by looking at the difference in reconstruction error. The results confirm that an autoencoder has the potential to be used as an aid in classifying exoplanet and non-exoplanet light curves.

FlashTalks / 23

Machine Learning Based Parametrization of Solar Active Regions Using Disentangled Variational Autoencoders

Speaker: Ekaterina Dineva (Centre for mathematical Plasma-Astrophysics)

Space Weather is an important aspect of our environment as a technologically dependent society. Moreover, space weather is a system of phenomena, bounded by complex cause-and-effect relations. In the century of Big Data solar physics, there are many archives of synoptic full-disk observations from space missions and ground-based facilities (which are growing daily) of multi-wavelength data. Driven by the amount of data and task, depending on data mining and dimensionality reduction, Machine Learning (ML) is rapidly integrated into solar and Heliophysics research. We use Deep Learning (DL) algorithms, i.e., Disentangled Variational Autoencoders (VAE, such as CT-VAE, betaVAE) for dimensionality reduction and feature extraction. The main goal of the VAE is to compress a high-dimensional input into a low-dimensional space, i.e., latent space (LS), reduce it down to its basic underlying structure, and retain enough information, that can be used to reconstruct the input and, if done properly, generate new data. Furthermore, disentangled representation learning facilitates the interpretability of the LS. The utilized dataset, Spaceweather HMI Active Region Patches (SHARPs) vector magnetic field (VMF) maps, is a by-product from the Solar Dynamics Observatory (SDO) and its Helioseismic and Magnetic Imager (HMI) full-disk observations. The Disentangled VAE are incorporated into a data pipeline, used for time series analysis, visualization and classification of the LS based on the SHARP VMF maps. The pipeline include unsupervised learning methods for label discovery and classification, such as T-SNE, UMAP, and HDBSCAN. Furthermore, the SHAPR ML-feature database will be used as a complementary dataset to the empirical SHARP parameters in a supervised DL space weather forecast pipeline.

FlashTalks / 24

GraphNeT: A Deep Learning Library for Neutrino Telescopes

Speaker: Rasmus Ørsøe (TUM)

GraphNeT, a deep learning framework for neutrino telescopes, provides a common library for deep learning experts and neutrino physicists in need of advanced deep learning techniques for their analyses. GraphNeT is a detector agnostic framework, making it easy and intuitive to apply deep learning techniques from one experiment to another, and to export models to

the physicists that relies on them. GraphNeT lowers technical barriers, enabling physicists to utilize deep learning without extensive expertise, and sets the stage for inter-experimental collaboration. This poster presentation highlights the key developments of GraphNeT 2.0 and their impact on how they provide a way for neutrino telescopes to work together on advancing the state-of-the-art for deep learning based techniques in the field.

FlashTalks / 25

Real-Time Detection of Cosmic Ray Systematic Glitches for the LiteBIRD satellite

Speaker: Emile Carinos (Université Toulouse 3 - IRAP)

The LiteBIRD space mission aims to measure the polarised Cosmic Microwave Background with unprecedented sensitivity with a satellite orbiting the L2 point, making it susceptible to cosmic ray effects. Impacts of cosmic rays on high-sensitivity bolometers can cause systematic effects such as glitches and their mitigation and substraction have been a key point of concern for past space missions such as the Planck telescope. For the LiteBIRD mission, due to the way the data will be sampled most of the cosmic ray glitches will not be visible in the transmitted data. In this work we will explore and discuss real-time methods to detect and possibly mitigate such glitches. We take inspiration from the work pursued at LHC to implement Neural Networks on FPGAs for fast real-time processing of large volumes of data.